https://doi.org/10.1038/s41588-018-0311-9

nature genetics

Corrected: Author Correction

Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk

Iris E. Jansen[1,2,48], Jeanne E. Savage[1,48], Kyoko Watanabe[1], Julien Bryois[3], Dylan M. Williams[3], Stacy Steinberg[4], Julia Sealock[5], Ida K. Karlsson[3,6], Sara Hägg[3], Lavinia Athanasiu[7,8], Nicola Voyle[9], Petroula Proitsi[9], Aree Witoelar[7,10], Sven Stringer[1], Dag Aarsland[9,11], Ina S. Almdahl[12,13,14], Fred Andersen[15], Sverre Bergh[16,17], Francesco Bettella[7,10], Sigurbjorn Bjornsson[18], Anne Brækhus[16,19], Geir Bråthen[20,21], Christiaan de Leeuw[1], Rahul S. Desikan[22], Srdjan Djurovic[7,23], Logan Dumitrescu[24,25], Tormod Fladby[12,13], Timothy J. Hohman[24,25], Palmi V. Jonsson[18,26], Steven J. Kiddle[27], Arvid Rongve[28,29], Ingvild Saltvedt[20,30], Sigrid B. Sando[20,21], Geir Selbæk[16,31], Maryam Shoai[32], Nathan G. Skene[33,34], Jon Snaedal[18], Eystein Stordal[35,36], Ingun D. Ulstein[37], Yunpeng Wang[7,10], Linda R. White[20,21], John Hardy[32], Jens Hjerling-Leffler[33], Patrick F. Sullivan[3,38,39], Wiesje M. van der Flier[2], Richard Dobson[9,40,41,42], Lea K. Davis[25,43], Hreinn Stefansson[4], Kari Stefansson[4], Nancy L. Pedersen[3], Stephan Ripke[44,45,46,49], Ole A. Andreassen[7,10,49] and Danielle Posthuma[1,47,49]*

Alzheimer's disease (AD) is highly heritable and recent studies have identified over 20 disease-associated genomic loci. Yet these only explain a small proportion of the genetic variance, indicating that undiscovered loci remain. Here, we performed a large genome-wide association study of clinically diagnosed AD and AD-by-proxy (71,880 cases, 383,378 controls). AD-by-proxy, based on parental diagnoses, showed strong genetic correlation with AD ($r_g = 0.81$). Meta-analysis identified 29 risk loci, implicating 215 potential causative genes. Associated genes are strongly expressed in immune-related tissues and cell types (spleen, liver, and microglia). Gene-set analyses indicate biological mechanisms involved in lipid-related processes and degradation of amyloid precursor proteins. We show strong genetic correlations with multiple health-related outcomes, and Mendelian randomization results suggest a protective effect of cognitive ability on AD risk. These results are a step forward in identifying the genetic factors that contribute to AD risk and add novel insights into the neurobiology of AD.

AD is the most frequent neurodegenerative disease with roughly 35 million people affected[1]. AD is highly heritable, with estimates ranging between 60% and 80%[2]. Genetically, AD can be roughly divided into two subgroups: (1) familial early-onset cases that are often explained by rare variants with a strong effect[3], and (2) late-onset cases that are influenced by multiple common variants with low effect sizes[4]. Segregation analyses have linked several genes to the first subgroup, including *APP*[5], *PSEN1*[6], and *PSEN2*[7]. The identification of these genes has resulted in valuable insights into a molecular mechanism with an important role in AD pathogenesis, the amyloidogenic pathway[8], exemplifying how gene discovery can add to biological understanding of disease etiology.

Besides the identification of a few rare genetic factors (for example *TREM2*[9] and *ABCA7*[10]), genome-wide association studies (GWASs) have mostly discovered common risk variants for the more complex late-onset type of AD. *APOE* is the strongest genetic risk locus for late-onset AD, responsible for a 3- to 15-fold increase in risk[11]. A total of 19 additional GWAS loci have been described using a discovery sample of 17,008 AD cases and 37,154 controls, followed by replication of the implicated loci with 8,572 AD patients and 11,312 controls[4]. The currently confirmed AD risk loci explain only a fraction of the heritability of AD and increasing the sample size is likely to boost the power for detection of more common risk variants, which will aid in understanding biological mechanisms involved in the risk for AD.

In the current study, we included 455,258 individuals (N_{sum}) of European ancestry, meta-analyzed in 3 phases (Fig. 1). Phase 1 consisted of 24,087 clinically diagnosed late-onset AD cases, paired with 55,058 controls. In phase 2, we analyzed an AD-by-proxy phenotype, based on individuals in the UK Biobank (UKB) for whom parental AD status was available (*N* proxy cases = 47,793; *N* proxy controls = 328,320). The value of by-proxy phenotypes for GWAS was recently demonstrated by Liu et al.[12] for 12 common diseases, including substantial gains in statistical power for AD. The high heritability of AD implies that case status for offspring can be partially inferred from parental case status and that offspring of AD parents are likely to have a higher genetic AD risk load. We thus defined individuals with one or two parents with AD as proxy cases,

A full list of affiliations appears at the end of the paper.

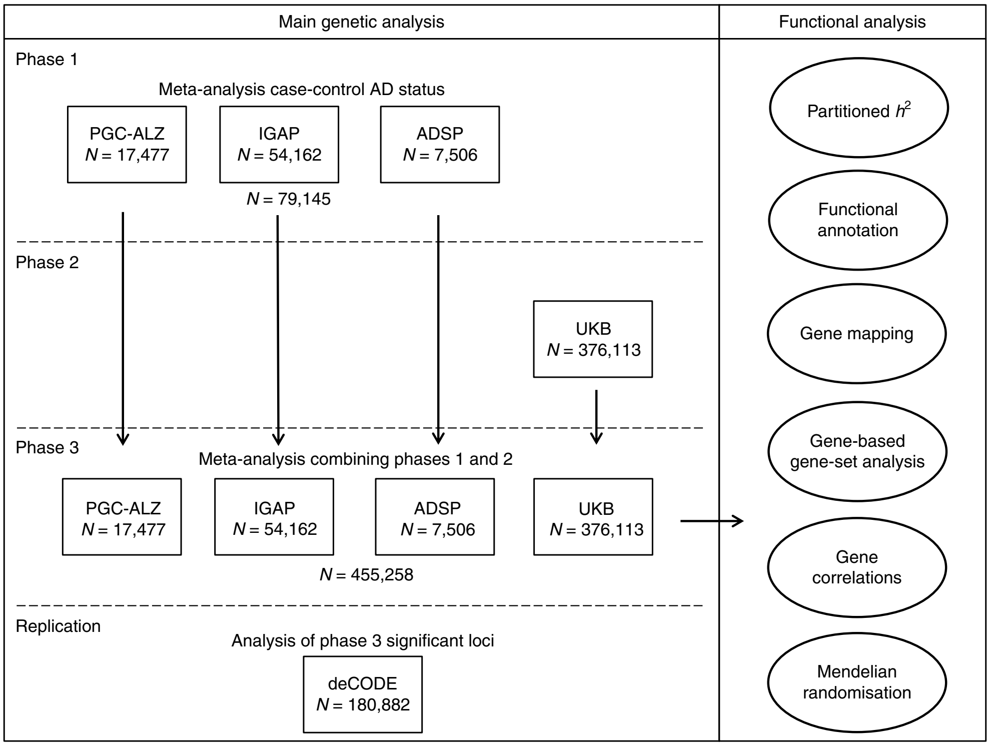

Fig. 1 | Overview of analysis steps. The main genetic analysis encompasses the procedures to detect GWAS risk loci for AD. The functional analysis includes the in silico functional follow-up procedures with the aim to put the genetic findings in biological context. N = total of individuals within specified data set.

while upweighting cases with two parents. Similarly, the proxy controls include subjects with two parents without AD, where older cognitively normal parents were upweighted to account for the higher likelihood that younger parents may still develop AD (see Methods). As the proxy phenotype is not a pure measure of an individual's AD status and may include individuals that never develop AD, genetic effect sizes will be somewhat underestimated. However, the proxy case-control sample is very large, and therefore substantially increases power to detect genetic effects for AD[12], as was also demonstrated in a more recent study using UKB[13]. Finally, in phase 3, we meta-analyzed all individuals of phase 1 and phase 2 together and tested for replication in an independent sample.

Results

Genome-wide meta-analysis for AD status. Phase 1 involved a genome-wide meta-analysis for clinically diagnosed AD case-control status using cohorts collected by 3 independent consortia (Alzheimer's disease working group of the Psychiatric Genomics Consortium (PGC-ALZ), the International Genomics of Alzheimer's Project (IGAP), and the Alzheimer's Disease Sequencing Project (ADSP)), totaling 79,145 individuals (N_{sum};effective sample size N_{eff} = 72,500) of European ancestry and 9,862,738 genetic variants passing quality control (Fig. 1 and Supplementary Table 1). The ADSP subset encompassed whole-exome sequencing data from 4,343 cases and 3,163 controls, while the remaining data sets consisted of genotype SNP arrays. For PGC-ALZ and ADSP, raw genotypic data were subjected to a standardized quality control pipeline. GWA analyses were run per cohort and then included in a meta-analysis alongside IGAP, for which only summary statistics were available (see Methods). As described in detail in the Supplementary Note, the phase 1 analysis identified 18 independent loci meeting genome-wide significance (GWS; $P < 5 \times 10^{-8}$), all of which have been identified by previous GWASs (Table 1, Supplementary Fig. 1, and Supplementary Table 2).

We next (phase 2) performed a GWAS using 376,113 individuals of European ancestry from UKB with parental AD status weighted by age to construct an AD-by-proxy status (Fig. 1). Here, we identified 13 independent GWS loci, 8 of which overlapped with phase 1 (Table 1 and Supplementary Note). We observed a strong genetic correlation of 0.81 (s.e.m. = 0.185) between AD status and AD-by-proxy, as well as substantial concordance in the individual SNP effects, as described in the Supplementary Note.

Given the high genetic overlap, in phase 3 we conducted a meta-analysis of the clinical AD GWASs and the AD-by-proxy GWAS (Fig. 1), comprising a total sample size of 455,258 (N_{eff} = 450,734), including 71,880 (proxy) cases and 383,378 (proxy) controls. The linkage disequilibrium (LD) score intercept[14] was 1.0018 (s.e.m. = 0.0109) and the sample size-adjusted[15] $\lambda_{1,000}$ was 1.044, indicating that most of the inflation in genetic signal (λ_{GC} = 1.0833) could be explained by polygenicity (Supplementary Fig. 1b). There were 2,357 GWS variants, which were represented by 94 lead SNPs, located in 29 distinct loci (Table 1, Fig. 2, and Supplementary Fig. 2). These included 15 of the 18 loci detected in phase 1, all of the 13 detected in phase 2, as well as 9 loci that were subthreshold in both individual analyses but reached significance in the meta-analysis. A large proportion of the lead SNPs (60 of 94) was concentrated in the established *APOE* risk locus on chromosome 19. This region is known to have a complex LD structure and a very strong effect on AD risk; thus, we consider these SNPs likely to represent a single association signal. Conditional analysis indicated that most loci represented a single fully independent signal, while the *TREM2*, *PTK2B/CLU*, and *APOE* loci contained multiple possible causal signals (Supplementary Note and Supplementary Tables 3 and 4).

Of the 29 associated loci, 16 overlapped 1 of the 20 genomic regions previously identified by the GWAS of Lambert et al.[4], replicating their findings, while 13 were novel. The association signals of five loci (*CR1*, *ZCWPW1*, *CLU/PTK2B*, *MS4A6a*, and *APH1B*) are partly based on the ADSP exome-sequencing data. Re-analysis

Table 1 | Summary statistics of significantly associated regions identified in the genome-wide association analysis of AD case-control status, AD-by-proxy phenotype, and meta-analysis

Region			Case-control status (phase 1)		AD-by-proxy (phase 2)		Overall (phase 3)							
Locus	Chr	Gene	SNP	*P*	SNP	*P*	SNP	bp	A1	A2	MAF	*Z*	*P*	Direction
1	1	***ADAMTS4***	rs4575098	1.57×10^{-4}	rs4575098	6.88×10^{-8}	rs4575098	161155392	A	G	0.240	6.36	**2.05×10^{-10}**	?+++
2	1	*CR1*	rs6656401	**1.39×10^{-17}**	rs679515	**8.85×10^{-10}**	rs2093760	207786828	A	G	0.205	8.82	**1.10×10^{-18}**	++++
3	2	*BIN1*	rs4663105	**3.58×10^{-29}**	rs4663105	**5.46×10^{-26}**	rs4663105	127891427	C	A	0.415	13.94	**3.38×10^{-44}**	?+++
4	2	*INPPD5*	rs10933431	1.67×10^{-6}	rs10933431	2.51×10^{-6}	rs10933431	233981912	G	C	0.235	−6.13	**8.92×10^{-10}**	?---
5	3	***HESX1***	NA		rs184384746	**1.24×10^{-8}**	rs184384746	57226150	T	C	0.002	5.69	**1.24×10^{-8}**	???+
6	4	***CLNK***	rs6448453	0.024	rs6448451	**1.19×10^{-8}**	rs6448453	11026028	A	G	0.252	6.00	**1.93×10^{-9}**	?+-+
--	4	*HS3ST1*	rs7657553	**2.16×10^{-8}**	rs7657553	0.790	rs7657553	11723235	A	G	0.291	1.95	0.051	?++-
7	6	*HLA-DRB1*	rs9269853	**2.66×10^{-8}**	rs6931277	1.78×10^{-7}	rs6931277	32583357	T	A	0.153	−6.49	**8.41×10^{-11}**	?---
8	6	*TREM2*	NA		rs187370608	**1.45×10^{-16}**	rs187370608	40942196	A	G	0.002	8.26	**1.45×10^{-16}**	???+
9	6	*CD2AP*	rs9381563	**5.35×10^{-9}**	rs9381563	8.10×10^{-6}	rs9381563	47432637	C	T	0.355	6.33	**2.52×10^{-10}**	?+++
10	7	*ZCWPW1*	rs1859788	**6.05×10^{-9}**	rs7384878	**2.38×10^{-10}**	rs1859788	99971834	A	G	0.310	−7.93	**2.22×10^{-15}**	----
11	7	*EPHA1*	rs11763230	**2.58×10^{-11}**	rs7810606	1.01×10^{-6}	rs7810606	143108158	T	C	0.500	−6.62	**3.59×10^{-11}**	?---
12	7	***CNTNAP2***	NA		rs114360492	**2.10×10^{-9}**	rs114360492	145950029	T	C	2.59×10^{-4}	5.99	**2.10×10^{-9}**	???+
13	8	*CLU/PTK2B*	rs4236673	**6.36×10^{-20}**	rs1532278	**7.45×10^{-9}**	rs4236673	27464929	A	G	0.391	−8.98	**2.61×10^{-19}**	----
14	10	*ECHDC3*	rs11257242	**2.38×10^{-8}**	rs11257238	5.84×10^{-5}	rs11257238	11717397	C	T	0.375	5.69	**1.26×10^{-8}**	?+++
15	11	*MS4A6A*	rs7935829	**8.21×10^{-13}**	rs1582763	**4.72×10^{-9}**	rs2081545	59958380	A	C	0.381	−7.97	**1.55×10^{-15}**	----
16	11	*PICALM*	rs10792832	**1.12×10^{-17}**	rs3844143	**5.31×10^{-11}**	rs867611	85776544	G	A	0.314	−8.75	**2.19×10^{-18}**	?---
17	11	*SORL1*	rs11218343	**5.57×10^{-11}**	rs11218343	2.81×10^{-6}	rs11218343	121435587	C	T	0.040	−6.79	**1.09×10^{-11}**	?---
18	14	*SLC24A4*	rs12590654	**1.98×10^{-8}**	rs12590654	3.70×10^{-6}	rs12590654	92938855	A	G	0.344	−6.39	**1.65×10^{-10}**	?---
19	15	***ADAM10***	rs442495	3.09×10^{-4}	rs442495	2.65×10^{-7}	rs442495	59022615	C	T	0.320	−6.07	**1.31×10^{-9}**	?---
20	15	***APH1B***	rs117618017	0.022	rs117618017	2.64×10^{-7}	rs117618017	63569902	T	C	0.132	5.52	**3.35×10^{-8}**	++++
21	16	***KAT8***	rs59735493	8.25×10^{-4}	rs59735493	3.72×10^{-6}	rs59735493	31133100	A	G	0.300	−5.49	**3.98×10^{-8}**	?---
22	17	*SCIMP*	rs113260531	3.21×10^{-6}	rs9916042	**4.73×10^{-8}**	rs113260531	5138980	A	G	0.120	6.12	**9.16×10^{-10}**	?+++
23	17	*ABI3*	rs28394864	7.29×10^{-5}	rs28394864	6.80×10^{-6}	rs28394864	47450775	A	G	0.473	5.62	**1.87×10^{-8}**	?+++
--	17	*BZRAP1-AS1*	rs2632516	**1.42×10^{-9}**	rs2632516	0.005	rs2632516	56409089	C	G	0.455	−4.90	9.66×10^{-7}	?---
--	18	*SUZ12P1*	rs8093731	**4.63×10^{-8}**	rs8093731	0.766	rs8093731	29088958	T	C	0.010	−2.17	0.030	?-?-
24	18	***ALPK2***	rs76726049	0.039	rs76726049	1.83×10^{-7}	rs76726049	56189459	C	T	0.014	5.52	**3.30×10^{-8}**	?+++
25	19	*ABCA7*	rs4147929	**8.64×10^{-9}**	rs3752241	**2.87×10^{-8}**	rs111278892	1039323	G	C	0.161	6.50	**7.93×10^{-11}**	?+++
26	19	*APOE*	rs41289512	**2.70×10^{-194}**	rs75627662	**9.51×10^{-296}**	rs41289512	45351516	G	C	0.039	35.50	**5.79×10^{-276}**	?+++
27	19	***AC074212.3***	rs76320948	1.54×10^{-5}	rs76320948	1.80×10^{-5}	rs76320948	46241841	T	C	0.046	5.46	**4.64×10^{-8}**	?+?+
28	19	*CD33*	rs3865444	**4.25×10^{-8}**	rs3865444	4.97×10^{-5}	rs3865444	51727962	A	C	0.320	−5.81	**6.34×10^{-9}**	?---
29	20	*CASS4*	rs6014724	8.72×10^{-8}	rs6014724	6.32×10^{-6}	rs6014724	54998544	G	A	0.089	−6.18	**6.56×10^{-10}**	?---

Note: independent lead SNPs are defined by r^2 <0.1; distinct genomic loci are >250 kb apart. The locus column indicates the loci number based on phase 3 (-- indicates that this locus is non-significant). The gene symbols are included to conveniently compare the significant loci with previously discovered loci. The bolded genes correspond to the novel loci indicating the genes in closest proximity to the most significant SNP, while emphasizing that this is not necessarily the causal gene. Allele1 is the effect allele for the meta-association statistic. The directions of effect of the distinct cohorts are in the following order: ADSP, IGAP, PGC-ALZ, and UKB; note that the first cohort is often missing as this concerns exome sequencing data. Corrected *P* value for significance equals 5×10^{-8} (marked as bold and underlined values). Note that the lead SNP can differ between the distinct analyses, while it tags the same locus.

of these loci excluding ADSP resulted in similar association signals (Supplementary Table 5), implying that we have correctly adjusted for partial sample overlap between IGAP and ADSP. The lead SNPs in 3 loci (with nearest genes *HESX1*, *TREM2*, and *CNTNAP2*) were only available in the UKB cohort (Table 1), but were of good quality (imputation quality INFO score >0.91, Hardy-Weinberg equilibrium *P* >0.19, missingness <0.003). These SNPs were all rare (minor allele frequency (MAF) <0.003), meaning that they will require future confirmation in another similarly large sample. However, variants in *TREM2* have been robustly linked to AD in previous research[9].

Verifying the 13 novel loci against other recent genetic studies on AD[9,12,16–18], 4 loci (*TREM2*, *ECHDC3*, *SCIMP*, and *ABI3*) have been previously discovered in addition to the 16 identified by Lambert et al., leaving 9 novel loci at the time of this writing (*ADAMTS4*, *HESX1*, *CLNK*, *CNTNAP2*, *ADAM10*, *APH1B*, *KAT8*, *ALPK2*, and *AC074212.3*). The *ADAMTS4* and *KAT8* loci have also since been identified in a recent analysis in a partially overlapping sample[13]. Comparing our meta-analysis results with all loci of Lambert et al.[4] to determine differences in associated loci, we were unable to observe 4 loci (*MEF2C*, *NME8*, *CELF1*, and *FERMT2*) at a GWS level (observed *P* values were 1.6×10^{-5} to 0.0011), which was mostly caused by a lower association signal in the UKB data set (Supplementary Table 6). By contrast, Lambert et al.[4] were unable to replicate the *DSG2* and *CD33* loci in the second stage of their study. In our study, *DSG2* was also not supported (meta-analysis $P=0.030$; UKB analysis $P=0.766$), implying invalidation of this locus, while the *CD33* locus (rs3865444 in Table 1) was significantly associated with AD (meta-analysis $P=6.34\times10^{-9}$; UKB analysis $P=4.97\times10^{-5}$), implying a genuine genetic association with AD risk.

Next, we aimed to find further support for the novel findings by using an independent Icelandic cohort (deCODE[19,20]), including 6,593 AD cases and 174,289 controls (Fig. 1 and Supplementary Table 7), to test replication of the lead SNP or an LD-proxy of the lead SNP (r^2 >0.9) in each locus. We were unable to test two loci

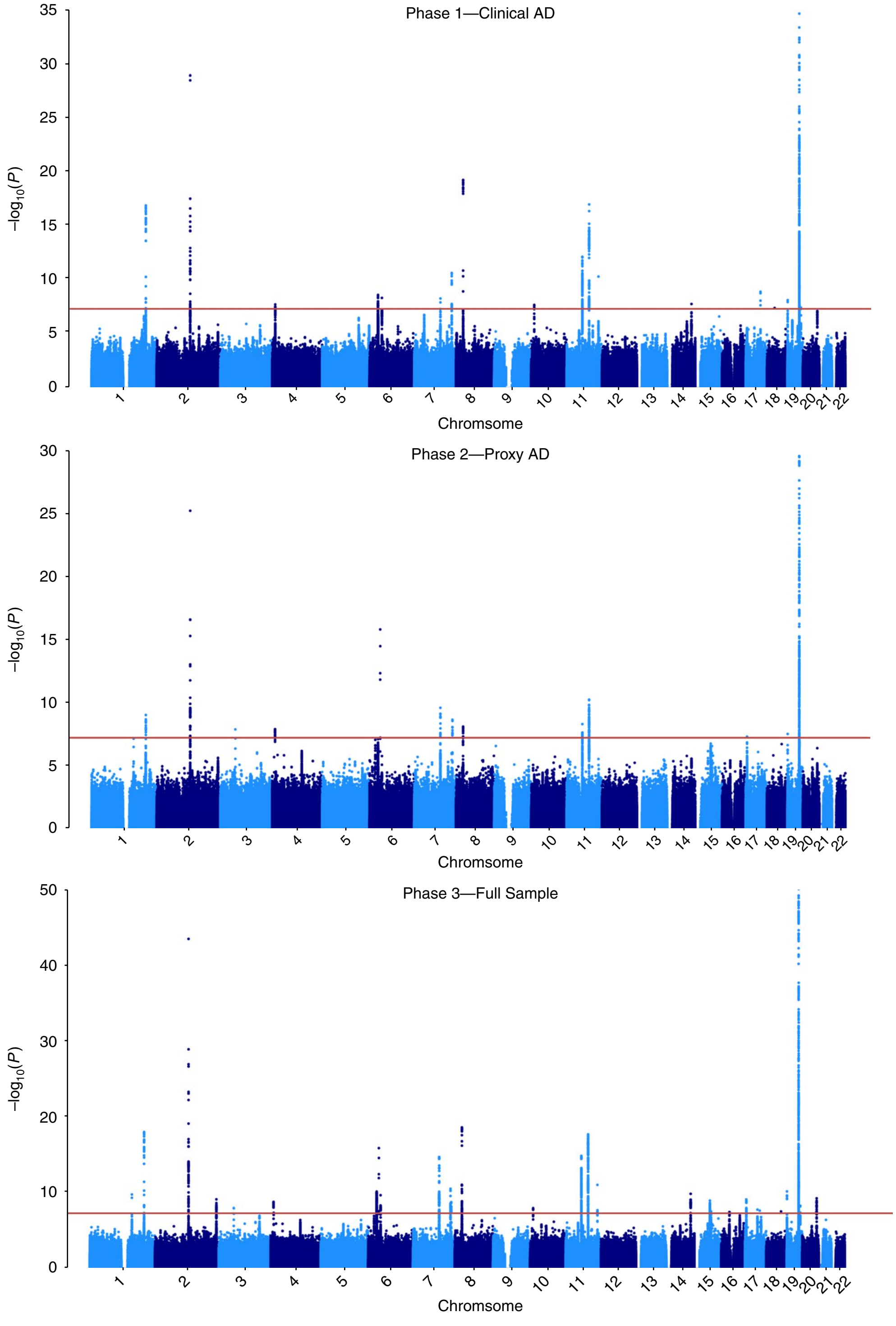

Fig. 2 | GWAS meta-analysis for AD risk (N = 455,258). Manhattan plot displays all associations per variant ordered according to their genomic position on the x axis and showing the strength of the association with the $-\log_{10}$-transformed P values on the y axis. The y axis is limited to enable visualization of non-*APOE* loci. For the phase 3 meta-analysis, the original $-\log_{10} P$ value for the APOE locus is 276.

as the lead SNPs (and SNPs in high LD) either were not present in the Icelandic reference panel or were not imputed with sufficient quality. For 6 of the 7 novel loci tested for replication, we observed the same direction of effect in the deCODE cohort. Furthermore, 4 loci (*CLNK*, *ADAM10*, *APH1B*, and *AC074212.3*) showed nominally significant association results (P <0.05) for the same SNP or an SNP in high LD (r^2 >0.9) within the same locus (two-tailed binomial test $P = 1.9 \times 10^{-4}$). The locus on chromosome 1 (*ADAMTS4*) was very close to significance ($P = 0.053$), implying stronger evidence for replication than for non-replication. Apart from the novel loci, we also observed sign concordance for 96.3% of the top (per-locus) lead SNPs in all loci from the meta-analysis (two-tailed binomial test $P = 4.17 \times 10^{-7}$) that were available in deCODE (26 of 27).

As an additional method of testing for replication, we used genome-wide polygenic score prediction in two independent samples[21]. The current results explain 7.1% of the variance in clinical AD at a low best-fitting P-threshold of 1.69×10^{-5} in 761 individuals with case-control diagnoses ($P=1.80\times10^{-10}$). When excluding the *APOE* locus (chr19: 45020859–45844508), the results explain 3.9% of the variance with a best-fitting P-threshold of 3.5×10^{-5} ($P=1.90\times10^{-6}$). We also predict AD status in a sample of 1,459 pathologically confirmed cases and controls[22] with an $R^2=0.41$ and an area under the curve (AUC) of 0.827 (95% confidence interval (95% CI): 0.805–0.849, $P=9.71\times10^{-70}$) using the best-fitting model of SNPs with a GWAS $P<0.50$, as well as $R^2=0.23$ and AUC = 0.733 (95% CI: 0.706–0.758, $P=1.16\times10^{-45}$) using only *APOE* SNPs. This validation sample contains a small number of individuals overlapping with IGAP; previous simulations with this sample have indicated that this overfitting increases the margin of error of the estimate approximately 2–3%[22]. This sample, however, represented severe, late-stage AD cases contrasted with supernormal controls, so the polygenic prediction may be higher than expected for typical case-control or population samples.

Functional interpretation of genetic variants. Functional annotation of all GWS SNPs ($n=2,357$) in the associated loci showed that SNPs were mostly located in intronic/intergenic areas, but also in regions that were enriched for chromatin states 4 and 5, implying effects on active transcription (Fig. 3 and Supplementary Table 8). Twenty-five GWS SNPs were exonic non-synonymous (Fig. 3a and Supplementary Table 9) with probably deleterious impacts on gene function. Converging evidence of strong association ($Z>|7|$) and a high observed probability of a deleterious variant effect (Combined Annotation Dependent Depletion (CADD)[23] score ≥30) was found for rs75932628 (*TREM2*), rs142412517 (*TOMM40*), and rs7412 (*APOE*). The first two missense mutations are rare (MAF = 0.002 and 0.001, respectively) and the alternative alleles were associated with higher risk for AD. The latter *APOE* missense mutation is the well-established protective allele Apoε2. Supplementary Tables 8 and 9 present a detailed annotation catalog of variants in the associated genomic loci. We also applied a fine-mapping model[24] to identify credible sets of causal SNPs from the identified GWS variants (Supplementary Table 8). The proportion of plausible causal SNPs varied drastically between loci; for example, 30 of 854 SNPs were selected in the *APOE* locus (no. 26), while 345 of 434 SNPs were nominated in the *HLA-DRB1* locus (no. 7). Credible causal SNPs were not limited to known functional categories such as exonic non-synonymous, indicating more complicated causal pathways that merit investigation with the set of variants prioritized by these statistical and functional annotations.

Partitioned heritability analysis[25], excluding SNPs with extremely large effect sizes (that is, *APOE* variants), showed enrichment for the SNP-heritability (h^2_{SNP}) for variants located in H3K27ac marks (enrichment = 3.18, $P=9.63\times10^{-5}$), which are associated with activation of transcription, and in super enhancers (enrichment = 3.62, $P=2.28\times10^{-4}$), which are genomic regions where multiple epigenetic marks of active transcription are clustered (Fig. 3d and Supplementary Table 10). Heritability was also enriched in variants on chromosome 17 (enrichment = 3.61, $P=1.63\times10^{-4}$) and we observed a trend of enrichment for heritability in common rather than rarer variants (Supplementary Fig. 3 and Supplementary Tables 11 and 12). Although a large proportion (23.9%) of the heritability can be explained by SNPs on chromosome 19, this enrichment is not significant due to the large standard errors around this estimate (Supplementary Table 11). Overall, these results suggest that, despite some non-synonymous variants contributing to AD risk, most of the GWS SNPs are located in non-coding regions and are enriched for regions that have an activating effect on transcription.

Implicated genes. To link the associated variants to genes, we applied three gene-mapping strategies implemented in Functional Mapping and Annotation (FUMA)[26] (see Methods). We used all SNPs with a P value $<5\times10^{-8}$ for gene mapping. Positional gene mapping aligned SNPs to 99 genes by their location within or immediately up/downstream (±10 kilobases (kb)) of known gene boundaries, expression quantitative trait locus (eQTL) gene mapping matched cis-eQTL SNPs to 168 genes whose expression levels they influence in 1 or more tissues, and chromatin interaction mapping linked SNPs to 21 genes based on three-dimensional DNA–DNA interactions between each SNP's genomic region and nearby or distant genes, which we limited to include only interactions between annotated enhancer and promoter regions (Supplementary Fig. 4 and Supplementary Tables 13 and 14). This resulted in 192 uniquely mapped genes, 80 of which were implicated by at least 2 mapping strategies and 16 by all 3 (Fig. 4e).

Of special interest is the locus on chromosome 8 (*CLU/PTK2B*). In the GWAS by Lambert et al.[4], this locus was defined as two distinct loci (*CLU* and *PTK2B*). Although our conditional analysis based on genetic data also specified this locus as having at least two independent association signals (Supplementary Table 4), the chromatin interaction data in two immune-related tissues—the spleen and liver (Supplementary Table 14)—suggest that the genomic regions indexed by *PTK2B* and *CLU* loci might physically interact (Fig. 3e), therefore putatively affecting AD pathogenesis via the same biological mechanism. The patterns of tissue-specific gene expression are largely dissimilar between *CLU* and *PTK2B*, although both are expressed relatively highly in the brain and lymph nodes[27]. Future studies should thus consider the joint effects of how these two genes simultaneously impact AD risk.

Eight genes (*HLA-DRB5*, *HLA-DRB1*, *HLA-DQA*, *HLA-DQB1*, *KAT8*, *PRSS36*, *ZNF232*, and *CEACAM19*) are particularly notable as they are implicated via eQTL association in the hippocampus, a brain region highly affected early in AD pathogenesis (Supplementary Table 13). Chromosome 16 contains a locus implicated by long-range eQTL association (Fig. 3f), clearly illustrating how the more distant genes *C16orf93*, *RNF40*, and *ITGAX* can be affected by a genetic factor (rs59735493) in various body tissues (for example, blood and skin), including a change in expression for *RNF40* observed in the dorsolateral prefrontal cortex. These observations emphasize the relevance of considering putative causal genes or regulatory elements not solely on the physical location but also on epigenetic influences. As detailed in the Supplementary Note, eQTLs were overrepresented in the risk loci and a number of quantitative trait locus (QTL) associations (including eQTLs, methylation quantitative trait loci (mQTLs), and histone acetylation quantitative trait loci (haQTLs)) were identified in relevant brain regions, providing interesting targets for future functional follow-up and biological interpretation (Supplementary Tables 15–17).

Although these gene-mapping strategies imply multiple putative causal genes per GWAS locus, several genes are of particular interest, as they have functional or previous genetic association with AD. For locus 1 in Supplementary Table 13, *ADAMTS4* encodes a protein of the ADAMTS family which has a function in neuroplasticity and has been extensively studied for its role in AD pathogenesis[28]. For locus 19, the obvious most likely causal gene is *ADAM10*, as this gene has been associated with AD by research focusing on rare coding variants in *ADAM10*[29]. Here, this gene is implicated as a common risk factor for AD, and this is supported by the putative causal molecular mechanism observed in dorsolateral prefrontal cortex eQTL and mQTL data (Supplementary Tables 15 and 16) for multiple common SNPs in LD. The lead SNP for locus 20 is a non-synonymous variant in exon 1 of *APH1B*, which encodes for a protein subunit of the γ-secretase complex cleaving APP[30]. A highly promising candidate gene for locus 21 is *KAT8*, as the lead SNP of this locus is located within the third intron of *KAT8*, and multiple significant variants within this locus influence the expression or methylation

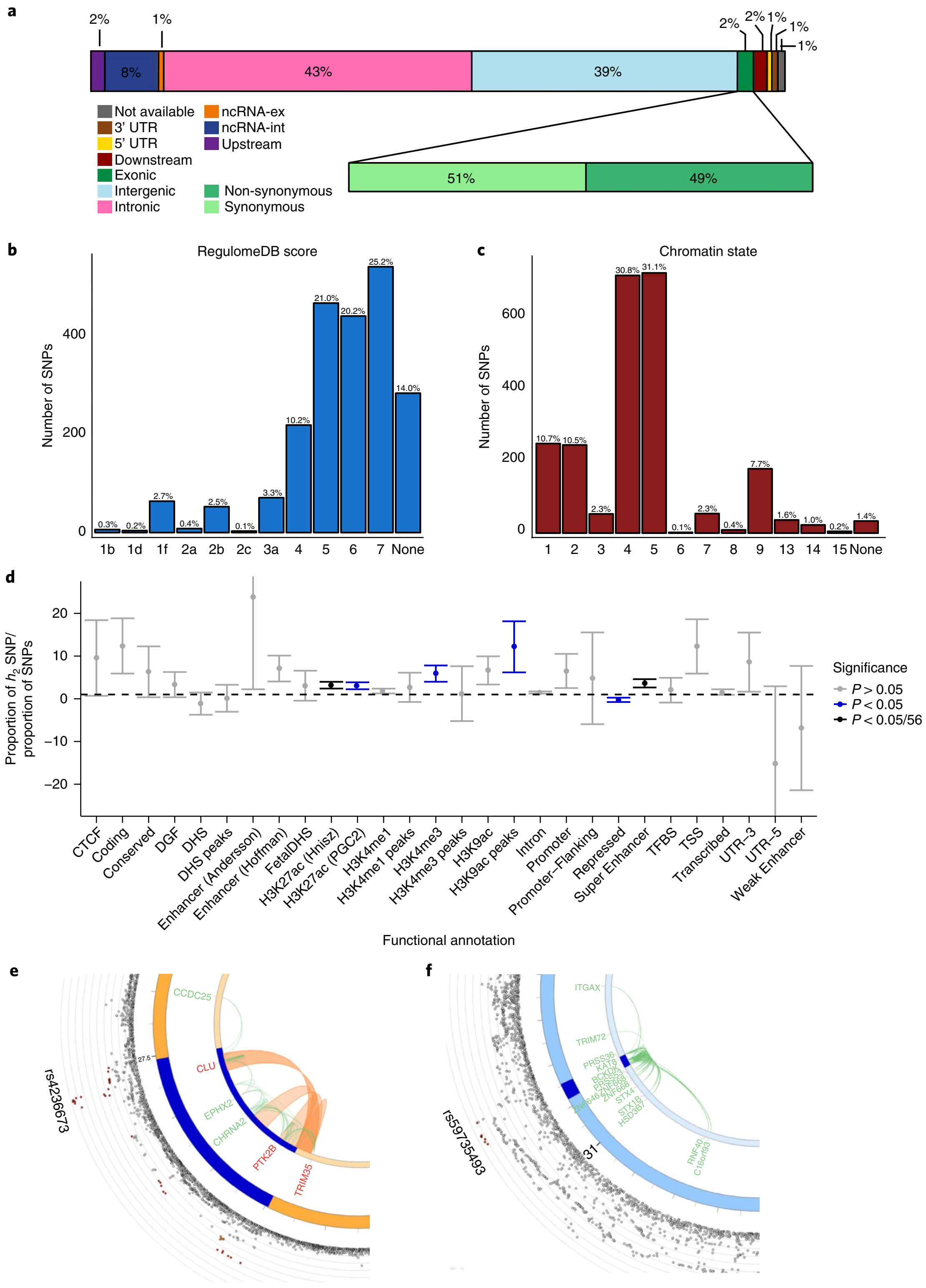

Fig. 3 | Functional annotation of GWAS results. a, Functional effects of variants in genomic risk loci of the meta-analysis (the colors of the legend are ordered from right to left in the figure)—the second bar shows distribution for exonic variants only. **b**, Distribution of RegulomeDB score for variants in genomic risk loci, with a low score indicating a higher probability of having a regulatory function (see Methods). **c**, Distribution of minimum chromatin state across 127 tissue and cell types for variants in genomic risk loci, with lower states indicating higher accessibility (see Methods). **d**, Heritability enrichment of 28 functional variant annotations calculated with stratified LD score regression (bars represent standard errors). CTCF, CCCTC-binding factor; DHS, DNaseI hypersensitive site; TFBS, transcription factor binding site; DGF, DNAaseI digital genomic footprint. **e**, Zoomed-in circos plot of chromosome 8. **f**, Zoomed-in circos plot of chromosome 16. Circos plots show implicated genes by significant loci, where dark blue areas indicate genomic risk loci, green lines indicate eQTL associations, and orange lines indicate chromatin interactions. Genes mapped by both eQTL and chromatin interactions are in red. The outer layer shows a Manhattan plot containing the negative $\log_{10}$-transformed P value of each SNP in the GWAS meta-analysis of AD. Full circos plots of all autosomal chromosomes are provided in Supplementary Fig. 4.

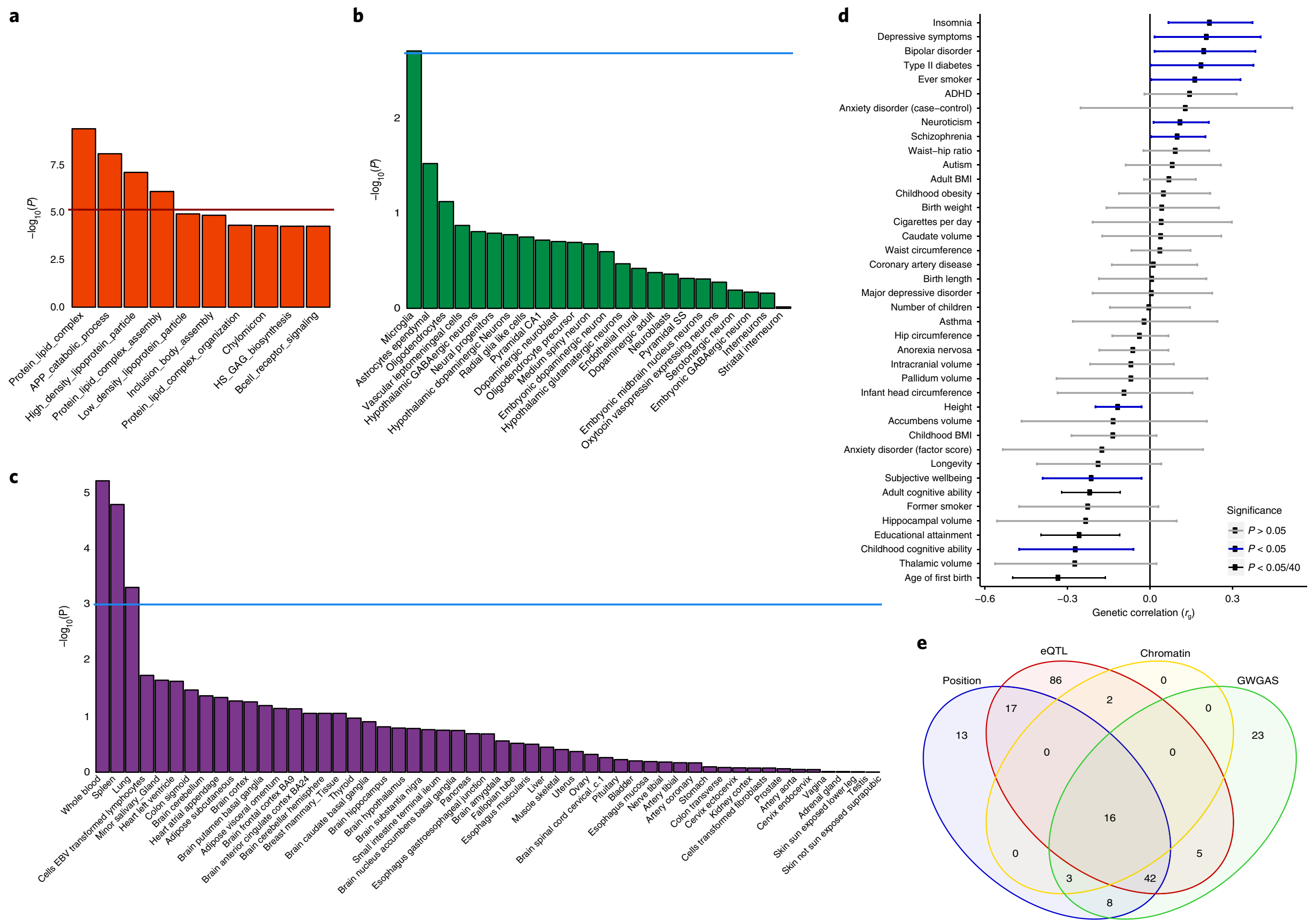

Fig. 4 | Functional implications based on gene-set analysis, genetic correlations, and functional annotations. a–c, The gene-set results are displayed per category of biological mechanisms (**a**), brain cell types (**b**), and tissue types (**c**). The red horizontal line indicates the significance threshold corrected for all gene-set tests of all categories, while the blue horizontal lines display the significance threshold corrected only for the number of tests within the three categories (that is, gene ontology, tissue expression, or single-cell expression). **d**, Genetic correlations between AD and other heritable traits (bars represent 95% confidence intervals). **e**, Venn diagram showing the number of genes mapped by four distinct strategies. ADHD, attention deficit hyperactivity disorder; BMI, body mass index; EBV, Epstein-Barr virus.

levels of *KAT8* in multiple brain regions (Supplementary Tables 13 and 16), including the hippocampus. The chromatin modifier *KAT8* is regulated by *KANSL1*, a gene associated with AD in absence of Apoε4. A study on Parkinson's disease reported KAT8 as a potential causal gene based on GWAS and differential gene expression results, implying a putative shared role in neurodegeneration of *KAT8* in AD and Parkinson's disease[31]. Although previously reported functional information on genes can be of great value, it is preferable to consider all implicated genes as putative causal factors to guide potential functional follow-up experiments.

We next performed genome-wide gene-based association analysis (GWGAS) using Multi-marker Analysis of GenoMic Annotation (MAGMA)[32]. This method annotates SNPs to known protein-coding genes to estimate aggregate associations based on all SNPs in a gene. It differs from FUMA as it provides a statistical gene-based test, whereas FUMA maps individually significant SNPs to genes. With GWGAS, we identified 97 genes that were significantly associated with AD (Supplementary Fig. 5 and Supplementary Table 18), of which 74 were also mapped by FUMA (Fig. 4e). In total, 16 genes were implicated by all 4 strategies (Supplementary Table 19), of which 7 genes (*HLA-DRA*, *HLA-DRB1*, *PTK2B*, *CLU*, *MS4A3*, *SCIMP*, and *RABEP1*) are not located in the *APOE* locus, and are therefore of high interest for further investigation.

Gene sets implicated in AD and AD-by-proxy. Using the gene-based *P* values, we performed gene-set analysis for curated biological pathways and tissue/single-cell expression. Four gene ontology[33] gene sets were significantly associated with AD risk: Protein lipid complex ($P=3.93\times10^{-10}$), Regulation of amyloid precursor protein catabolic process ($P=8.16\times10^{-9}$), High density lipoprotein particle ($P=7.81\times10^{-8}$), and Protein lipid complex assembly ($P=7.96\times10^{-7}$) (Fig. 4a and Supplementary Tables 20 and 21). Conditional analysis on the *APOE* locus showed associations with AD for these four gene sets to be independent of the effect of *APOE*, though part of the association signal was also attributable to *APOE*. All 25 genes of the High density lipoprotein particle pathway are also part of the Protein lipid complex; conditional analysis showed that these gene sets are not interpretable as independent associations ($P=0.18$), but the other three sets are independently significant (Supplementary Table 20).

Linking gene-based *P* values to tissue- and cell type-specific gene sets, no association survived the stringent Bonferroni correction, which corrected for all tested gene sets (that is, 6,994 gene ontology categories, 53 tissues, and 39 cell types). However, we did observe suggestive associations across immune-related tissues when correcting only for the number of tests within all tissue types or cell types (Fig. 4c and Supplementary Table 22), particularly whole blood

($P=5.61\times10^{-6}$), spleen ($P=1.50\times10^{-5}$), and lung ($P=4.67\times10^{-4}$), which were independent from the *APOE* locus. In brain single-cell expression gene-set analyses, we found association for microglia in the mouse-based expression data set ($P=1.96\times10^{-3}$), though not surviving the stringent Bonferroni correction (Fig. 4b and Supplementary Table 23). However, we observed a similar association signal for microglia in a second independent single-cell expression data set in humans ($P=2.56\times10^{-3}$) (Supplementary Fig. 6 and Supplementary Table 24). As anticipated, both microglia signals are partly depending on *APOE*, though a large part is independent (Supplementary Tables 23 and 24).

Cross-trait genetic influences. As described in the Supplementary Note and Supplementary Tables 25 and 26, we observed that the genetic influences on AD overlapped with a number of other diseases and psychological traits including cognitive ability and educational attainment, replicating previous studies[34,35]. To extend these findings, we used generalized summary statistic-based Mendelian randomisation[36] (GSMR) to test for potential credible causal associations of genetically correlated outcomes which may directly influence the risk for AD. Due to the nature of AD being a late-onset disorder and summary statistics for most other traits being obtained from younger samples, we do not report tests for the opposite direction of potential causality (that is, we did not test for a causal effect of a late-onset disease on an early-onset disease). In this set of analyses, SNPs from the summary statistics of genetically correlated phenotypes were used as instrumental variables to estimate the putative causal effect of these 'exposure' phenotypes on AD risk by comparing the ratio of SNPs' associations with each exposure to their associations with AD outcome (see Methods). Association statistics were standardized, such that the reported effects reflect the expected difference in odds ratio for AD as a function of every standard deviation increase in the exposure phenotype. We observed a protective effect of cognitive ability (odds ratio = 0.89, 95% CI: 0.85–0.92, $P=5.07\times10^{-9}$), educational attainment (odds ratio = 0.88, 95% CI: 0.81–0.94, $P=3.94\times10^{-4}$), and height (odds ratio = 0.96, 95% CI: 0.94–0.97, $P=1.84\times10^{-8}$) on risk for AD (Supplementary Table 27 and Supplementary Fig. 7). No substantial evidence of pleiotropy was observed between AD and these phenotypes, with <1% of overlapping SNPs being filtered as outliers (Supplementary Table 27).

Discussion

By using an unconventional approach of including a proxy phenotype for AD to increase sample size, we have identified nine novel loci and gained novel biological knowledge on AD etiology. We were able to test seven of the nine novel loci for replication, of which four loci showed clear replication, one locus showed marginal replication, and two loci were not replicated at this moment. Both the high genetic correlation between the standard case-control status and the UKB by-proxy phenotype ($r_g=0.81$) and the high rate of novel locus replication in the independent deCODE cohort suggest that this strategy is robust. Through in silico functional follow-up analysis, and in line with previous research[18,37], we emphasize the crucial causal role of the immune system—rather than immune response as a consequence of disease pathology—by establishing variant enrichments for immune-related body tissues (whole blood, spleen, and liver) and for the main immune cells of the brain (microglia). Of note, the enrichment observed for liver could alternatively indicate the genetic involvement of the lipid system in AD pathogenesis[38]. Furthermore, we observe informative eQTL associations and chromatin interactions within immune-related tissues for the identified genomic risk loci. Together with the AD-associated genetic effects on lipid metabolism in our study, these biological implications (which are based on genetic signals and unbiased by earlier biological beliefs) strengthen the hypothesis that AD pathogenesis involves an interplay between inflammation and lipids, as lipid changes might harm immune responses of microglia and astrocytes, and vascular health of the brain[39].

In accordance with previous clinical research, our study suggests an important role for protective effects of several human traits on AD. Cognitive reserve has been proposed as a protective mechanism in which the brain aims to control brain damage with previously existing cognitive processing strategies[40]. Our findings imply that some component of the genetic factors for AD might affect cognitive reserve, rather than being involved in AD pathology-related damaging processes, influencing AD pathogenesis in an indirect way through cognitive reserve. Furthermore, a large-scale community-based study observed that AD incidence rates declined over decades, which was specific for individuals with at minimum a high school diploma[41]. Combined with our Mendelian randomization results for educational attainment, this suggests that the protective effect of educational attainment on AD is influenced by genetics. Similarly, the observed positive effects of height could be a result of the genetic overlap between height and intracranial volume[42,43], a measure associated with decreased risk of AD[44]. This indirect association is furthermore supported by the observed increase in cognitive reserve for taller individuals[45]. Alternatively, genetic variants influencing height might also affect biological mechanisms involved in AD etiology, such as *IGF1* that codes for insulin-like growth factor and is associated with cerebral amyloid[46].

The results of this study could furthermore serve as a valuable resource for selection of promising genes for functional follow-up experiments and identifying targets for drug development and stratification approaches. We anticipate that functional interpretation strategies and follow-up experiments will result in a comprehensive understanding of late-onset AD etiology, which will serve as a solid foundation for improvement of AD therapy.

URLs. UK Biobank, http://ukbiobank.ac.uk; Database of Genotypes and Phenotypes (dbGaP), https://www.ncbi.nlm.nih.gov/gap; FUMA software, http://fuma.ctglab.nl; MAGMA software, http://ctg.cncr.nl/software/magma; mvGWAMA and effective sample size calculation, https://github.com/Kyoko-wtnb/mvGWAMA; LD Score Regression software, https://github.com/bulik/ldsc; LD Hub (GWAS summary statistics), http://ldsc.broadinstitute.org/; LD scores, https://data.broadinstitute.org/alkesgroup/LDSCORE/; Psychiatric Genomics Consortium (GWAS summary statistics), http://www.med.unc.edu/pgc/results-and-downloads; MSigDB curated gene-set database, http://software.broadinstitute.org/gsea/msigdb/collections.jsp; NHGRI GWAS catalog, https://www.ebi.ac.uk/gwas/; GSMR software, http://cnsgenomics.com/software/gsmr/; credible SNP set analysis software, https://github.com/hailianghuang/FM-summary.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0311-9.

Received: 1 March 2018; Accepted: 9 November 2018;
Published online: 7 January 2019

References

1. Prince, M. et al. The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimers Dement.* **9**, 63–75.e2 (2013).
2. Gatz, M. et al. Role of genes and environments for explaining Alzheimer disease. *Arch. Gen. Psychiatry* **63**, 168–174 (2006).
3. Cacace, R., Sleegers, K. & Van Broeckhoven, C. Molecular genetics of early-onset Alzheimer's disease revisited. *Alzheimers Dement.* **12**, 733–748 (2016).
4. 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).

5. Goate, A. et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **349**, 704–706 (1991).
6. Sherrington, R. et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375**, 754–760 (1995).
7. Sherrington, R. et al. Alzheimer's disease associated with mutations in presenilin 2 is rare and variably penetrant. *Hum. Mol. Genet.* **5**, 985–988 (1996).
8. Karran, E., Mercken, M. & De Strooper, B. The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nat. Rev. Drug Discov.* **10**, 698–712 (2011).
9. Jonsson, T. et al. Variant of TREM2 associated with the risk of Alzheimer's disease. *N. Engl. J. Med.* **368**, 107–116 (2013).
10. Steinberg, S. et al. Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. *Nat. Genet.* **47**, 445–447 (2015).
11. Liu, C. C., Liu, C. C., Kanekiyo, T., Xu, H. & Bu, G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat. Rev. Neurol.* **9**, 106–118 (2013).
12. 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).
13. Marioni, R. E. et al. GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 99 (2018).
14. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
15. de Bakker, P. I. W. et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–R128 (2008).
16. Guerreiro, R. et al. TREM2 variants in Alzheimer's disease. *N. Engl. J. Med.* **368**, 117–127 (2013).
17. Desikan, R. S. et al. Polygenic overlap between C-reactive protein, plasma lipids, and Alzheimer disease. *Circulation* **131**, 2061–2069 (2015).
18. 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).
19. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
20. Steinthorsdottir, V. et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298 (2014).
21. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: polygenic risk score software. *Bioinformatics* **31**, 1466–1468 (2015).
22. Valentina, E. P., J., M. A., Matt, H. & John, H. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. *Ann. Neurol.* **82**, 311–314 (2017).
23. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
24. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
25. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
26. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
27. Fagerberg, L. et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13**, 397–406 (2014).
28. Gurses, M. S., Ural, M. N., Gulec, M. A., Akyol, O. & Akyol, S. Pathophysiological function of ADAMTS enzymes on molecular mechanism of Alzheimer's disease. *Aging Dis.* **7**, 479–490 (2016).
29. Suh, J. et al. ADAM10 missense mutations potentiate beta-amyloid accumulation by impairing prodomain chaperone function. *Neuron* **80**, 385–401 (2013).
30. Dries, D. R. & Yu, G. Assembly, maturation, and trafficking of the gamma-secretase complex in Alzheimer's disease. *Curr. Alzheimer Res.* **5**, 132–146 (2008).
31. Dumitriu, A. et al. Integrative analyses of proteomics and RNA transcriptomics implicate mitochondrial processes, protein folding pathways and GWAS loci in Parkinson disease. *BMC Med. Genomics* **9**, 5 (2016).
32. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
33. The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).
34. Anttila, V. et al. Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).
35. Savage, J. E. et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
36. Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
37. Skene, N. G. & Grant, S. G. Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Front. Neurosci.* **10**, 16 (2016).
38. Kang, J. & Rivest, S. Lipid metabolism and neuroinflammation in Alzheimer's disease: a role for liver X receptors. *Endocr. Rev.* **33**, 715–746 (2012).
39. Loewendorf, A., Fonteh, A., Mg, H. & Me, C. Inflammation in Alzheimer's disease: cross-talk between lipids and innate immune cells of the brain. *J. Immun. Res.* **2**, 1022 (2015).
40. Stern, Y. Cognitive reserve in ageing and Alzheimer's disease. *Lancet Neurol.* **11**, 1006–1012 (2012).
41. Satizabal, C., Beiser, A. S. & Seshadri, S. Incidence of dementia over three decades in the Framingham Heart Study. *N. Engl. J. Med.* **375**, 93–94 (2016).
42. Adams, H. H. et al. Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nat. Neurosci.* **19**, 1569–1582 (2016).
43. Ikram, M. A. et al. Common variants at 6q22 and 17q21 are associated with intracranial volume. *Nat. Genet.* **44**, 539–544 (2012).
44. Graves, A. B. et al. Head circumference as a measure of cognitive reserve. Association with severity of impairment in Alzheimer's disease. *Br. J. Psychiatry* **169**, 86–92 (1996).
45. Abbott, R. D. et al. Height as a marker of childhood development and late-life cognitive function: the Honolulu-Asia Aging Study. *Pediatrics* **102**, 602–609 (1998).
46. Giuffrida, M. L. et al. Beta-amyloid monomer and insulin/IGF-1 signaling in Alzheimer's disease. *Mol. Neurobiol.* **46**, 605–613 (2012).

Acknowledgements

This work was funded by the Netherlands Organization for Scientific Research (NWO VICI 453-14-005). The analyses were carried out on the Genetic Cluster Computer, which is financed by the Netherlands Scientific Organization (NWO: 480-05-003), by the VU University, Amsterdam, the Netherlands, and by the Dutch Brain Foundation, and is hosted by the Dutch National Computing and Networking Services SurfSARA. The work was also funded by the Research Council of Norway (grant nos. 251134, 248778, 223273, 213837, and 225989), KG Jebsen Stiftelsen, the Norwegian Health Association, European Community's JPND Program, ApGeM RCN grant no. 237250, and the European Community's grant no. PIAPP-GA-2011-286213 PsychDPC. This research has been conducted using the UK Biobank resource under application number 16406 and the public ADSP data set, obtained through the Database of Genotypes and Phenotypes under accession number phs000572. Full acknowledgments for the studies that contributed data can be found in the Supplementary Note. We thank the numerous participants, researchers, and staff from many studies who collected and contributed to the data.

Author contributions

I.E.J. and J.E.S. performed the analyses. D.P. and O.A.A. conceived the idea for the study. D.P. and S.R. supervised analyses. Sv.St. performed quality control on the UK Biobank data and wrote the analysis pipeline. K.W. constructed and applied the FUMA pipeline for performing follow-up analyses. J.B. conducted the single-cell enrichment analyses. J.H.L. and N.S. contributed data. M.S. and J.H. performed polygenic score analyses. D.P. and I.E.J. wrote the first draft of the paper. All other authors contributed data and critically reviewed the paper.

Competing interests

The authors report the following potentially competing financial interests. P.F.S.: Lundbeck (advisory committee), Pfizer (Scientific Advisory Board member), and Roche (grant recipient, speaker reimbursement). J.H.L.: Cartana (Scientific Advisor) and Roche (grant recipient). O.A.A.: Lundbeck (speaker's honorarium). St.St., H.S., and K.S. are employees of deCODE Genetics/Amgen. J.H. is a cograntee of Cytox from Innovate UK (UK Department of Business). D.A. has received research support and/or honoraria from Astra-Zeneca, Lundbeck, Novartis Pharmaceuticals, and GE Health, and serves as a paid consultant for Lundbeck, Eisai, Heptares, and Axovant. All other authors declare no financial interests or potential conflicts of interest.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0311-9.

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

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

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

[1]Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University, Amsterdam, the Netherlands. [2]Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, the Netherlands. [3]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [4]deCODE Genetics/Amgen, Reykjavik, Iceland. [5]Interdisciplinary Graduate Program, Vanderbilt University, Nashville, TN, USA. [6]Institute of Gerontology and Aging Research Network–Jönköping (ARN-J), School of Health and Welfare, Jönköping University, Jönköping, Sweden. [7]NORMENT, K.G. Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [8]Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway. [9]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [10]Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [11]Center for Age-Related Diseases, Stavanger University Hospital, Stavanger, Norway. [12]Department of Neurology, Akershus University Hospital, Lørenskog, Norway. [13]AHUS Campus, University of Oslo, Oslo, Norway. [14]Department of Psychiatry of Old Age, Oslo University Hospital, Oslo, Norway. [15]Department of Community Medicine, University of Tromsø, Tromsø, Norway. [16]Norwegian National Advisory Unit on Ageing and Health, Vestfold Hospital Trust, Tønsberg, Norway. [17]Centre for Old Age Psychiatry Research, Innlandet Hospital Trust, Ottestad, Norway. [18]Department of Geriatric Medicine, Landspitali University Hospital, Reykjavik, Iceland. [19]Geriatric Department, Oslo University Hospital, Oslo, Norway. [20]Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway. [21]Department of Neurology, St Olav's Hospital, Trondheim University Hospital, Trondheim, Norway. [22]Neuroradiology Section, Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, USA. [23]Department of Medical Genetics, Oslo University Hospital, Oslo, Norway. [24]Vanderbilt Memory & Alzheimer's Center, Department of Neurology, Vanderbilt University Medical Center, Nashville, TN, USA. [25]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA. [26]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [27]MRC Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge, Cambridge, UK. [28]Department of Research and Innovation, Helse Fonna, Haugesund, Norway. [29]Department of Clinical Medicine, University of Bergen, Bergen, Norway. [30]Department of Geriatrics, St. Olav's Hospital, Trondheim University Hospital, Trondheim, Norway. [31]Institute of Health and Society, University of Oslo, Oslo, Norway. [32]Department of Neurodegenerative Disorders, Institute of Neurology, UCL, London, UK. [33]Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. [34]Department of Neuromuscular Diseases, Institute of Neurology, University College London, London, UK. [35]Department of Psychiatry, Namsos Hospital, Namsos, Norway. [36]Department of Mental Health, Norwegian University of Science and Technology, Trondheim, Norway. [37]Memory Clinic, Geriatric Department, Oslo University Hospital, Oslo, Norway. [38]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [39]Department of Psychiatry, University of North Carolina, Chapel Hill, NC, USA. [40]NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, UK. [41]Institute of Health Informatics Research, University College London, London, UK. [42]Health Data Research UK London, University College London, London, UK. [43]Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [44]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [45]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [46]Department of Psychiatry and Psychotherapy, Charité–Universitätsmedizin, Berlin, Germany. [47]Department of Clinical Genetics, VU University Medical Center, Amsterdam, the Netherlands. [48]These authors contributed equally: I.E. Jansen, J.E. Savage. [49]These authors jointly supervised this work: S. Ripke, O.A. Andreassen, D. Posthuma. *e-mail: d.posthuma@vu.nl

Methods

Participants. Participants in this study were obtained from multiple sources, including raw data from case-control samples collected by PGC-ALZ and ADSP (made publicly available through dbGaP (see URLs)), summary data from the case-control samples in the IGAP, and raw data from the population-based UKB sample which was used to create a weighted AD-by-proxy phenotype. An additional independent case-control sample (deCODE) was used for replication. Full descriptions of the samples and their respective phenotyping and genotyping procedures are provided in the Supplementary Note and the Life Sciences Reporting Summary.

Data analysis. *Single-marker association analysis.* Genome-wide association analysis for each of the ADSP, PGC-ALZ, and UKB data sets was performed in PLINK[47], using logistic regression for dichotomous phenotypes (cases versus controls for ADSP and PGC-ALZ cohorts), and linear regression for phenotypes analyzed as continuous outcomes (proxy phenotype constructed as the number of parents with AD for UKB cohort). For the ADSP and PGC-ALZ cohorts, association tests were adjusted for sex, batch (if applicable), and the first four ancestry principal components. Twenty principal components were calculated, and depending on the data set being tested, additional principal components (on top of the standard of four) were added if significantly associated with the phenotype. Furthermore, for the PGC-ALZ cohorts age was included as a covariate. For 4,537 controls of the DemGene cohort (subset of PGC-ALZ), no detailed age information was available, besides the age range the subjects were in (20–45 years). We therefore set the age of these individuals conservatively to 20 years. For the ADSP data set, age was not included as a covariate due to the enrichment for older controls (mean age cases = 73.1 years (s.e.m. = 7.8); mean age controls = 86.1 years (s.e.m. = 4.5)) in their collection procedures. Correcting for age in ADSP would remove a substantial part of genuine association signals (for example, well-established *APOE* locus rs11556505 is strongly associated with AD ($P = 1.08 \times 10^{-99}$), which is lost when correcting for age ($P = 0.0054$). For the UKB data set, 12 ancestry principal components were included as covariates, as well as age, sex, genotyping array, and assessment center. We used the genome-wide threshold for significance of $P < 5 \times 10^{-8}$.

Multivariate genome-wide meta-analysis. Two meta-analyses were performed, including: phase 1, cohorts with case-control phenotypes (IGAP, ADSP, and PGC-ALZ data sets); and phase 3, all cohorts, also including the UKB proxy phenotype.

Because of partial overlap between cohorts, the per SNP test statistic was defined by

$$Z_k = \frac{\sum_i w_i Z_i}{\sqrt{\sum_i w_i^2 + \sum_i \sum_j w_i w_j \ \mid CTI_{ij} \ \mid (i \neq j)}}$$

where w_i and Z_i are the squared root of the sample size and the test statistic of SNP k in cohort i, respectively. CTI is the cross-trait LD score intercept estimated by LD Score Regression (LDSC)[14,48] using genome-wide summary statistics. This is equal to[48]

$$CTI_{ij} = \frac{N_{sij} \ \rho_{ij}}{\sqrt{N_i \ N_j}}$$

where N_i and N_j are the sample sizes of cohorts i and j and N_{sij} is the number of samples overlapping between them, and ρ_{ij} is the phenotypic correlation between the measures used in the two cohorts for the overlapping samples. Under the null hypothesis of no association, any correlation between Z_i and Z_j is determined only by that phenotypic correlation, scaled by the relative degree of overlap. As such, this correlation can be estimated by the CTI.

The test statistics per SNP per GWAS were converted from the P value, incorporating the sign of either beta or odds ratio. When direction is aligned the conversion is two-sided. To avoid infinite values, we replaced P value 1 with 0.999999 and $P < 1 \times 10^{-323}$ with 1×10^{-323} (the minimum >0 value in Python). The script for the multivariate GWAS is available online (see URLs).

Effective sample size. The effective sample size (N_{eff}) is computed for each SNP k from the matrix M, containing the sample size N_i of each cohort i on the diagonal and the estimated number of shared data points $N_{sij}\rho_{ij} = CTI_{ij}\sqrt{N_i N_j}$ for each pair of cohorts i and j as the off-diagonal values. A recursive approach is used to compute N_{eff}. Going from the first cohort to the last, the (remaining) size of the current cohort is added to the total N_{eff}. Then, for each remaining other cohort it overlaps with, the size of the other cohort is reduced by the expected number of samples shared by the current cohort; overlap between the remaining cohorts is similarly adjusted. This process ensures that each overlapping data point is counted only once in N_{eff}.

The computation proceeds as follows. Starting with the first cohort in M, N_{eff} is first increased by $M_{1,1}$, corresponding to the sample size of that cohort. The proportion of samples shared between cohort 1 and each other cohort j is then computed as $p_{1,j} = M_{1,j}/M_{j,j}$, and M is adjusted to remove this overlap, multiplying all values in each column j by $1 - p_{1,j}$. This amounts to reducing the sample size of each other cohort j by the number of samples it shares with cohort 1 and reducing the shared samples between cohort j and subsequent cohorts by the same proportion. After this, the first row and column of M are discarded, and the same process is applied to the new M matrix. This is repeated until M is empty.

The effective sample size is used as a parameter in the MAGMA analysis (see Methods "Gene based-analysis") and reported in the main text as the combined sample sizes for the meta-analysis. We use the term N_{sum} to indicate the total number of individuals when simply summing them over the distinct cohorts. The script for the N_{eff} computation is available online (see URLs).

Genomic risk loci definition. We used FUMA[26] v1.2.8, an online platform for functional mapping and annotation of genetic variants, to define genomic risk loci and obtain functional information of relevant SNPs in these loci. We first identified independent significant SNPs that have a genome-wide significant P value ($<5 \times 10^{-8}$) and are independent from each other at $r^2 < 0.6$. These SNPs were further represented by lead SNPs, which are a subset of the independent significant SNPs that are in approximate linkage equilibrium with each other at $r^2 > 0.6$. We then defined associated genomic risk loci by merging any physically overlapping lead SNPs (LD blocks <250 kb apart). LD information was calculated using the UKB genotype data as a reference.

For GWS SNPs in the defined risk loci, we applied a summary statistic-based fine-mapping model to identify credible causal SNPs within each locus, as previously described[24]. This Bayesian model estimates a per-SNP posterior probability of a true disease association using maximum likelihood estimation and the steepest descent approach, creating a set of SNPs in each locus that contains the causal SNP in 99% of cases, given that the causal variants are among the genotyped/imputed SNPs. The software used, FM-summary, is available online (see URLs).

Independent sample replication. For novel SNPs identified in the phase 3 meta-analysis, replication was tested in the independent deCODE sample using logistic regression with AD status as the response and genotype counts and a set of nuisance variables, including sex, county of birth, and current age, as predictors[20]. Correction for inflation of test statistics due to relatedness and population stratification in this Icelandic cohort was performed using the intercept estimate (1.29) from LD score regression[14].

Conditional analysis. We performed conditional analysis with GCTA-COJO[49] to assess the independence of association signals, either within or between GWAS risk loci. COJO enables conditional analysis of GWAS summary statistics without individual-level genotype data. We therefore performed conditional analysis on the phase 3 summary statistics, using 10,000 randomly selected unrelated samples from the UKB data set as a reference data set to determine LD patterns. Conditional analysis was run per chromosome or per locus with the default settings of the software.

Heritability and genetic correlation. LD score regression[14] was used to estimate clinical AD heritability and to calculate genetic correlations[48] between the case-control and proxy phenotypes using summary statistics. Precalculated LD scores from the 1000 Genomes European reference population were obtained online (see URLs). Liability heritability was calculated with a population prevalence of 0.043[1] (the population prevalence of age group 70–75 in the Western European population, resembling the average age of onset of 74.5 for the clinical case group) and a sample prevalence of 0.304. The genetic correlation was calculated on HapMap3 SNPs only to ensure high-quality LD score calculation.

Stratified heritability. To test whether specific categories of SNP annotations were enriched for heritability, we partitioned the SNP heritability for binary annotations using stratified LD score regression[14]. Heritability enrichment was calculated as the proportion of heritability explained by an SNP category divided by the proportion of SNPs that are in that category. Partitioned heritability was computed by 28 functional annotation categories, by MAF in 6 percentile bins, and by 22 chromosomes, resulting in a total of 56 statistical tests. Annotations for binary categories of functional genomic characteristics (for example, coding or regulatory regions) were obtained online (see URLs). The Bonferroni-corrected significance threshold for 56 annotations was set at $P < 0.05/56 = 8.93 \times 10^{-4}$.

Polygenic risk scoring. We calculated polygenic scores (PGSs) using two independent genotype data sets. First, 761 individuals (379 cases and 382 controls) from the ADDNeuroMed study[50] were included, using the same quality control and imputation approach as for the other data sets with genotype-level data (see Supplementary Note). Second, 1,459 individuals (912 severe, late-stage cases and 547 age-matched controls with little to no cognitive dysfunction) from the TGEN study[22] were assessed and their diagnostic status was confirmed via postmortem neuropathology. Imputed SNPs in this sample were filtered based on imputation-quality INFO score >0.9 and MAF >0.01. PGSs were created using PLINK[47] for the TGEN data set and PRSice[21] for the ADDNeuroMed data set. In both samples, PGSs were calculated on hard-called imputed genotypes using P value thresholds from 0.0 to 0.5 and using PLINK's clumping procedure to prune for LD. Clumping was based on the effect size estimates of SNPs originating from the phase 3 meta-analysis for the ADDNeuroMed sample. For TGEN, clumping was previously performed using the IGAP summary statistics; these clumped SNPs were filtered for overlap with the phase 3 SNPs. PGSs were calculated in both samples using the SNP effect size estimates from the phase 3 meta-analysis. The explained variance (ΔR^2) was derived from a linear model in which the AD phenotype was regressed on each PGS while controlling for GWAS covariates, compared to a linear model with covariates only. In the TGEN data set, sensitivity, specificity, and AUC for predicting confirmed case/control status were calculated, using the R package pROC[51] and bootstrapped confidence intervals. Of note, approximately 3% of the TGEN sample overlapped with the IGAP cohort included in the meta-analysis; previous simulation work using PGSs in this sample has shown that this overfitting leads to only a modest increase (2–3%) in the margin of error around the AUC estimate[22].

Functional annotation. Functional annotation of GWS SNPs implicated in the meta-analysis was performed using FUMA[26] v1.2.8. Functional consequences for these SNPs were obtained by matching SNPs to databases containing known functional annotations, including ANNOVAR[52] categories, CADD scores[23], RegulomeDB[53] scores, and chromatin states[54,55]. ANNOVAR annotates the functional consequences of SNPs on genes (for example, intron, exon, and intergenic). CADD scores predict how deleterious the effect of an SNP will be, with higher scores referring to higher deleteriousness. A CADD score above 12.37 is the threshold to be potentially pathogenic[56]. The RegulomeDB score is a categorical score based on information from eQTLs and chromatin marks, ranging from 1a to 7, with lower scores indicating an increased likelihood of having a regulatory function. The chromatin state represents the accessibility of genomic regions (every 200 base pairs (bp)), with 15 categorical states predicted by a hidden Markov model based on 5 chromatin marks in the Roadmap Epigenomics Project[55]. A lower state indicates higher accessibility, with states 1–7 referring to open chromatin states. We annotated the minimum chromatin state across tissues to SNPs. A legend describing the RegulomeDB and chromatin state scores can be found in the Supplementary Note.

Gene mapping. Genome-wide significant loci obtained by GWAS were mapped to genes in FUMA[26] using three strategies:

1. Positional mapping maps SNPs to genes based on physical distance (within a 10-kb window) from known protein-coding genes in the human reference assembly (GRCh37/hg19).
2. eQTL mapping maps SNPs to genes with which they show a significant eQTL association (that is, allelic variation at the SNP is associated with the expression level of that gene). eQTL mapping uses information from 45 tissue types in 3 data repositories (GTEx[57] v6, Blood eQTL browser[58], BIOS QTL browser[59]), and is based on cis-eQTLs which can map SNPs to genes up to 1 Mb apart. We used a false discovery rate of 0.05 to define significant eQTL associations.
3. Chromatin interaction mapping was performed to map SNPs to genes when there was a three-dimensional DNA–DNA interaction between the SNP region and another gene region. Chromatin interaction mapping can involve long-range interactions as it does not have a distance boundary. FUMA currently contains data on the 3D structure of chromatin interactions (Hi-C data) of 14 tissue types from the study of Schmitt et al.[60] Since chromatin interactions are often defined in a certain resolution, such as 40 kb, an interacting region can span multiple genes. If an SNP is located in a region that interacts with a region containing multiple genes, it will be mapped to each of those genes. To further prioritize candidate genes, we selected only genes mapped by chromatin interaction in which one region involved in the interaction overlaps with a predicted enhancer region in any of the 111 tissue/cell types from the Roadmap Epigenomics Project[55] and the other region is located in a gene promoter region (250 bp up- and 500 bp downstream of the transcription start site and also predicted by Roadmap to be a promoter region). This method reduces the number of genes mapped but increases the likelihood that those identified will have a plausible biological function. We used a false discovery rate of 1×10^{-5} to define significant interactions, based on previous recommendations[60] modified to account for the differences in cell lines used here.

Brain-specific QTL annotation. As AD is characterized by neurodegeneration, we annotated the significant genomic loci with publicly available databases of expression, methylation, and histone acetylation QTLs, as cataloged in BRAINEAC[61], CommonMind Consortium Portal[62], and xQTL Serve[63], as an extension of the GTEx tissue eQTL mapping performed in FUMA. Descriptions of these brain eQTL databases and the settings we used are in the Supplementary Note.

Gene-based analysis. To account for the distinct types of genetic data in this study, genotype array (PGC-ALZ, IGAP, UKB) and whole-exome sequencing data (ADSP), we first performed two GWGASs using MAGMA[32], followed by a meta-analysis. SNP-based *P* values from the meta-analysis of the three genotype array-based data sets were used as input for the first GWGAS, while the unimputed individual-level sequence data of ADSP were used as input for the second GWGAS. A total of 18,233 protein-coding genes (each containing at least one SNP in the GWAS) from the National Center for Biotechnology Information (NCBI) 37.3 gene definitions were used as a basis for GWGAS in MAGMA. Bonferroni correction was applied to correct for multiple testing ($P < 2.74\times10^{-6}$).

Gene-set analysis. Results from the GWGAS analyses were used to test for association in 7,086 predefined gene sets of four categories:

1. Curated gene sets (6,994) representing known biological and metabolic pathways derived from gene ontology (5917 gene sets), Biocarta (217 gene sets), KEGG (186 gene sets), and Reactome (674 gene sets), cataloged by and obtained from MsigDB version 6.1[64] (see URLs).
2. Gene expression values from 53 tissues obtained from GTEx[57], $\log_2$ transformed with pseudocount 1 after winsorization at 50 and averaged per tissue.
3. Cell type-specific expression in 24 broad categories of brain cell types, which were calculated following the method described in ref. [37]. Briefly, brain cell type expression data were drawn from single-cell RNA sequencing data from mouse brains. For each gene, the value for each cell type was calculated by dividing the mean unique molecular identifier counts for the given cell type by the summed mean unique molecular identifier counts across all cell types. Single-cell gene sets were derived by grouping genes into 40 equal bins based on specificity of expression.
4. Nucleus-specific gene expression of 15 distinct human brain cell types from the study described in ref. [65]. The value for each cell type was calculated as in point (3).

These gene sets were tested using MAGMA. We computed competitive *P* values, which represent the test of association for a specific gene set compared with genes not in the gene set to correct for baseline level of genetic association in the data. The Bonferroni-corrected significance threshold accounts for the total number of statistical gene-set tests and was set to 0.05/7,087 gene sets $= 7.06\times10^{-6}$. The suggestive significance threshold was defined by the number of tests within the category. Conditional analyses were performed as a follow-up using MAGMA to test whether each significant association observed was independent of *APOE* (a gene set including all genes within region chr19: 45,020,859–45,844,508). Furthermore, the association between each of the significant gene sets was tested conditional on each of the other significantly associated gene sets. Gene sets that retained their association after correcting for other sets were considered to represent independent signals. We note that this is not a test of association per se, but rather a strategy to identify, among gene sets with known significant associations and overlap in genes, which set(s) are responsible for driving the observed association.

Cross-trait genetic correlation. Genetic correlations (r_g) between AD and 41 phenotypes were computed using LD score regression[14], based on GWAS summary statistics obtained from publicly available databases (see URLs and Supplementary Table 26). The Bonferroni-corrected significance threshold accounted for the total number of genetic correlation estimations and was set to 0.05/41 traits $= 1.22\times10^{-3}$.

Mendelian randomization. To infer credible causal associations between AD and traits that are genetically correlated with AD, we performed GSMR[36] (see URLs). This method utilizes summary-level data to test for putative causal associations between a risk factor (exposure) and an outcome by using independent genome-wide significant SNPs as instrumental variables as an index of the exposure. HEIDI outlier detection was used to filter genetic instruments that showed clear pleiotropic effects on the exposure phenotype and the outcome phenotype. We used a threshold *P* value of 0.01 for the outlier detection analysis in HEIDI, which removes 1% of SNPs by chance if there is no pleiotropic effect. To test for a potential causal effect of various outcomes on risk for AD, we selected phenotypes in non-overlapping samples that showed (suggestive) significant ($P < 0.05$) genetic correlations (r_g) with AD. With this method it is typical to test for bidirectional causation by repeating the analyses while switching the role of the exposure and the outcome; however, because AD is a late-onset disease, it makes little sense to estimate its causal effect on outcomes that develop earlier in life, particularly when the summary statistics for these outcomes were derived mostly from younger samples than those of AD cases. Therefore, we conducted these analyses only in one direction. For genetically correlated phenotypes, we selected independent ($r^2 \leq 0.1$), GWS lead SNPs as instrumental variables in the analyses. The method estimates a putative causal effect of the exposure on the outcome (b_{xy}) as a function of the relationship between the SNPs' effects on the exposure (b_{zx}) and the SNPs' effects on the outcome (b_{zy}), given the assumption that the effect of non-pleiotropic SNPs on an exposure (x) should be related to their effect on the outcome (y) in an independent sample only via mediation through the phenotypic causal pathway (b_{xy}). The estimated causal effect coefficients (b_{xy}) are approximately equal to the natural log odds ratio[36] for a case-control trait. An odds ratio of 2 can be interpreted as a doubled risk compared with the population prevalence of a binary trait for every standard deviation increase in the exposure trait. This method can help differentiate the causal direction of association between two traits, but cannot make any statement about the intermediate mechanisms involved in any potential causal process.

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

Code availability. The analyses were produced with standard code for software programs utilized, which can be made available from the corresponding author on reasonable request. All software used is freely available online. Custom code for the meta-analysis correcting for overlapping samples is available at https://github.com/Kyoko-wtnb/mvGWAMA.

Data availability

Summary statistics will be made available for download upon publication (https://ctg.cncr.nl/software/summary_statistics).

References

47. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
48. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
49. 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).
50. Lovestone, S. et al. AddNeuroMed—the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Ann. N. Y. Acad. Sci.* **1180**, 36–46 (2009).
51. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
52. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
53. Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
54. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
55. Roadmap Epigenomics Consortium, Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
56. Amendola, L. M. et al. Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Res.* **25**, 305–315 (2015).
57. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
58. Westra, H. J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
59. Zhernakova, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
60. Schmitt, A. D. et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17**, 2042–2059 (2016).
61. Ramasamy, A. et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.* **17**, 1418–1428 (2014).
62. Fromer, M. et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
63. 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).
64. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
65. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).

nature research

Corresponding author(s): Danielle Posthuma

Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
□	☒	An indication of 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 statistics 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
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

No software was used for data collection.

Data analysis

We used standard, publicly available statistical genetics software packages, which are described and linked to in the Online Methods. The packages we used included:

R - Data management and statistical analyses (R Core Team, 2016)
PLINK - Genetic association testing (Purcell et al., 2007; Chang et al., 2015)
METAL - GWAS meta-analysis (Willer et al., 2010)
FUMA - Online platform for functional annotation of GWAS results (Watanabe et al., 2017)
MAGMA - Gene-based association testing (de Leeuw et al., 2015)
LD score regression - SNP-based heritability and genetic correlations from GWAS summary statistics (Bulik-Sullivan et al., 2015)
PRSice - Polygenic score analysis (Eusden et al., 2015)
LDpred - Polygenic score analysis (Vilhjalmsson et al., 2015)
GSMR - Mendelian randomization (Zhu et al., 2018)
Credible SNP set analysis (Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2014)

Custom code for the meta-analysis correcting for overlapping samples is available at https://github.com/Kyoko-wtnb/mvGWAMA.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. 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 will be made available for download upon publication (https://ctg.cncr.nl).

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size

Data used in this study was collected by external sources and used for secondary analysis. Sample size was not pre-determined and was chosen based on all known available cohorts with relevant data collected to date, after quality control steps were performed in each cohort (described in detail in Online Methods). The sample size was calculated as the number of individuals summed across all studies in the meta-analysis, N=455,258 (also reported separately by individual study). Because there was minor overlap in individuals between samples, we calculated the effective sample size of independent observations (Neff=450,734), as described in the methods. Power calculations using the Genetic Power Calculator indicated that we had virtually 100% power to detect SNPs accounting for >0.01% of trait variance in our sample size.

Data exclusions

In case-control samples, controls for Alzheimer's disease were excluded for neurocognitive issues and young age, as described in detail in Online Methods. For genetic quality control, samples were excluded for poor quality DNA and genetic variants were excluded using standard metrics like minor allele frequency and imputation quality.

Replication

The meta-analysis strategy includes replication by default, as it weights the reported test statistics by the evidence of association across multiple samples. Further, SNP-based replication was carried out for the top GWAS association signals in an independent sample. Aggregate genomic associations were replicated using polygenic score validation in two separate independent samples.

Randomization

Not relevant; no experimental procedures were performed as this was a study of genetic association between genotypes and non-manipulated phenotypes.

Blinding

Not relevant; there were no experimental groups.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	□ Unique biological materials
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
□	☒ Human research participants

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Human adult research participants were included, with an overrepresentation of older individuals and females in the case-control samples due to the nature of Alzheimer's as a late-onset disease. Participants were collected from external studies with the aims of comparing AD cases and controls (PGC-ALZ, IGAP, ADSP) or investigating general health conditions in middle-aged adults (UK Biobank).
Recruitment	Participants from case-control studies were primarily recruited from clinics, nursing homes, disease registries, and hospitals, with controls being drawn from various ongoing studies and screened to exclude dementia/cognitive decline. The UK Biobank recruited adult volunteers from national health registration records. UK Biobank participants are healthier than the general population, but since the data used in this study referred to parental diagnoses, the impact of selection bias should be minor.