Genetic meta-analysis identifies 9 novel loci and functional pathways for Alzheimer's disease risk

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

1. Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University, Amsterdam, The Netherlands.
2. Alzheimer Center and Department of Neurology, Amsterdam Neuroscience, VU University Medical Center, 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, USA.
6. NORMENT, K.G. Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway.
7. Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway.
8. Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK.
9. Institute of Clinical Medicine, University of Oslo, Oslo, Norway
10. Center for Age-Related Diseases, Stavanger University Hospital, Stavanger, Norway.
11. Department of Neurology, Akershus University Hospital, Lørenskog, Norway.
12. AHUS Campus, University of Oslo, Oslo, Norway.
13. Department of Psychiatry of Old Age, Oslo University Hospital, Oslo, Norway.
14. Department of Community Medicine, University of Tromsø, Tromsø, Norway.
15. Norwegian National Advisory Unit on Ageing and Health, Vestfold Hospital Trust, Tønsberg, Norway.
16. Centre for Old Age Psychiatry Research, Innlandet Hospital Trust, Ottestad, Norway.
17. Department of Geriatric Medicine, Landspitali University Hospital, Reykjavik, Iceland.
18. Geriatric Department, University Hospital Oslo and University of Oslo, Oslo, Norway.
19. Department of Neuroscience, Norwegian University of Science and Technology, Trondheim, Norway.
20. Department of Neurology, St Olav's Hospital, Trondheim University Hospital, Trondheim, Norway.
21. Neuroradiology Section, Department of Radiology and Biomedical Imaging, University of California, San Francisco, USA.

22. Department of Medical Genetics, Oslo University Hospital, Oslo, Norway.
23. Vanderbilt Memory & Alzheimer's Center, Department of Neurology, Vanderbilt University Medical Center, Nashville, USA.
24. Faculty of Medicine, University of Iceland, Reykjavik, Iceland.
25. MRC Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge, Cambridge, UK.
26. Department of Research and Innovation, Helse Fonna, Oslo, Norway.
27. Department of Clinical Medicine, University of Bergen, Bergen, Norway.
28. Department of Geriatrics, St. Olav's Hospital, Trondheim University Hospital, Trondheim, Norway.
29. Institute of Health and Society, University of Oslo, Oslo, Norway.
30. Department of Molecular Neuroscience, Institute of Neurology, UCL London, United Kingdom
31. Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden.
32. Department of Psychiatry, Namsos Hospital, Namsos, Norway.
33. Department of Mental Health, Norwegian University of Science and Technology, Trondheim, Norway.
34. Memory Clinic, Geriatric Department, Oslo University Hospital, Oslo, Norway.
35. Department of Genetics, University of North Carolina, Chapel Hill, USA.
36. Department of Psychiatry, University of North Carolina, Chapel Hill, USA.
37. Farr Institute of Health Informatics Research, University College London, London, UK.
38. Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, US.
39. Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, US.
40. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, USA.
41. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, USA.
42. Dept. of Psychiatry and Psychotherapy, Charité - Universitätsmedizin, Berlin, Germany.
43. Department of Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands.

* These authors contributed equally to this work

#Correspondence to: Danielle Posthuma: Department of Complex Trait Genetics, VU University, De Boelelaan 1085, 1081 HV, Amsterdam, The Netherlands. Phone: +31 20 598 2823, Fax: +31 20 5986926, d.posthuma@vu.nl

Word count: Introductory paragraph: 209; main text: 5,363; Online methods: 6,170
Display items: 5 (Figures 4)
Includes **Supplementary Figures 1-7, Supplementary Tables 1-27.**

Abstract

Late onset Alzheimer's disease (AD) is the most common form of dementia with more than 35 million people affected worldwide, and no curative treatment available. AD is highly heritable and recent genome-wide meta-analyses have identified over 20 genomic loci associated with AD. Yet these only explain a small proportion of the genetic variance, indicating that undiscovered loci exist. Here, we performed the largest genome-wide association study of clinically diagnosed AD and AD-by-proxy (71,880 AD cases, 383,378 controls). AD-by-proxy status is based on parental AD diagnosis and showed strong genetic correlation with AD (r_g=0.81). Genetic meta-analysis identified 29 risk loci, of which 9 are novel, and implicating 215 potential causative genes. Independent replication further supports these novel loci in AD. Associated genes are strongly expressed in immune-related tissues and cell types (spleen, liver and microglia). Furthermore, gene-set analyses indicate the genetic contribution of 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 randomisation results suggest a protective effect of cognitive ability on AD risk. These results are a step forward in identifying more of the genetic factors that contribute to AD risk and add novel insights into the neurobiology of AD to guide new drug development.

Main text

Alzheimer's disease (AD) is the most frequent neurodegenerative disease with roughly 35 million people affected.[1] Results from twin studies indicate that AD is highly heritable, with estimates ranging between 60 and 80%.[2] Genetically, AD can be roughly divided into 2 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 aetiology.

Besides the identification of a few rare genetic factors (e.g. *TREM2*[9] and *ABCA7*[10]), genome-wide association studies (GWAS) 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, where heterozygous and homozygous Apoe ε4 carriers are predisposed for a 3-fold and 15-fold increase in risk, respectively.[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 more than 20 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-analysed in 3 phases (**Figure 1**). Phase 1 consisted of 24,087 clinically diagnosed late-onset AD cases, paired with 55,058 controls. In phase 2, we analysed an AD-by-proxy phenotype, based on individuals in the UK Biobank (UKB) for whom parental AD status was available (N proxy cases=74,793; N proxy controls=328,320; **Online Methods**). The value of the usage of by-proxy phenotypes for GWAS was recently demonstrated by Liu et al[12] for 12 common diseases. In particular for AD, Liu et al[12] report substantial gains in statistical power by using a proxy phenotype, based on simulations and confirmed using empirical data from the 1st release of UKB. We here applied the proxy phenotype strategy for AD in the UKB 2nd release (full sample). In this sample, parental diagnosis for AD was available for N=376,113 individuals, of whom 393 individuals had a known diagnosis of AD themselves (identified from medical register data). 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 (N=47,793), while upweighting the proxy cases with 2 parents. Similarly, the proxy controls include subjects with 2 parents without AD (N=328,320), where older cognitively normal parents were upweighted as proxy controls 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] We first analysed the clinically defined case-control samples separately from the proxy case-control sample to allow investigation of overlap in genetic signals for these two measurements of AD risk. Finally, in phase 3, we meta-analysed all individuals of phase 1 and phase 2 together and tested for replication in an independent sample.

Genome-wide meta-analysis for AD status

Phase 1 involved a genome-wide meta-analysis for AD case-control status using cohorts collected as part of 3 independent main consortia (PGC-ALZ, IGAP and ADSP), totalling 79,145 individuals (effective sample size N_{eff}=72,500) of European ancestry and 9,862,738 genetic variants passing quality control **(Figure 1, Supplementary Table 1**). The ADSP consortium obtained whole exome sequencing data from 4,343 cases and 3,163 controls, while the remaining datasets consisted of genotype single nucleotide polymorphism (SNP) array data. AD patients were diagnosed according to generally acknowledged diagnostic criteria, such as the NINCDS-ADRDA (see **Methods**). All cohorts for which we had access to the raw genotypic data were subjected to a standardized quality control pipeline, and GWA analyses were run per cohort and then included in a meta-analysis, alongside one dataset (IGAP) for which only summary statistics were available (see **Methods**). The full sample liability SNP-heritability (h^2_{SNP}), estimated with the linkage disequilibrium (LD) Score regression (LDSC) method, was 0.055 (SE=0.0099), implying that 5.5% of AD heritability can be explained by the tested SNPs. This is in line with previous estimates for IGAP (6.8%) also estimated by LDSC regression method, which is based on summary statistics.[13,14]

The λ_{GC}=1.10 indicated the presence of modestly inflated genetic signal compared to the null hypothesis of no association. The LD score intercept[14] was 1.044 (SE=0.0084) and the sample size-adjusted[15] λ_{1000} was 1.086, indicating that most inflation could be explained by polygenic signal (**Supplementary Figure 1**). In the meta-analysis of AD case-control status, 1,067 variants indexed by 51 lead SNPs in approximate linkage equilibrium (r^2<0.1) reached genome-wide significance (GWS; $P<5\times10^{-8}$) (**Supplementary Figure 1**; **Supplementary Table 2**). These were located in 18 distinct genomic loci (**Table 1**). 15 of these loci confirmed previous findings (Lambert et al[4]) in a sample partially overlapping with that of the current study. The 3 remaining loci

(*HS3ST1*, *ECHDC3* and *BZRAP1-AS1*[*]) have been linked more recently to AD in a genetic study[16] of AD-related cholesterol levels while conditioning on lipid levels and in a transethnic genome-wide association study of AD.[17]

We next (phase 2) performed a GWAS for AD-by-proxy using 376,113 individuals of European ancestry from the UKB version 2 release using parental AD status weighted by age to construct an AD-by-proxy status (**Figure 1;** see **Methods**). The LD score intercept was 1.022 (SE=0.0099) and the λ_{1000} was 1.032, indicating that most of the inflation in genetic signal (λ_{GC}=1.071) could be explained by polygenicity (**Supplementary Figure 1B**). For AD-by-proxy, 719 GWS variants were indexed by 61 lead SNPs in approximate linkage equilibrium (r^2<0.1), located in 13 loci (**Supplementary Figure 1A**). Of these, 8 loci overlapped with the significantly associated loci identified in phase 1 for clinical AD case control status (**Table 1**).

We observed a strong genetic correlation of 0.81 (SE= 0.185, using LDSC) between AD status and AD-by-proxy, indicating substantial overlap between genetic effects beyond shared GWS SNPs. Sign concordance tests indicated that 50.4% of all LD-independent (r^2<0.1) genome-wide SNPs (significant and non-significant) had consistent direction of effects between the two phenotypes (N=344,581 overlapping SNPs), slightly greater than the chance expectation of 50% (exact binomial test P=2.45×10^{-7}). Of the 18 strongest lead SNPs (one per locus) identified by the case-control meta-analysis, all were available in UKB and 94.4% were sign-concordant (P=1.45x10^{-4}), while of the 13 strongest lead SNPs (one per locus) identified in UKB, 10 were

[*] Although the gene that is in closest proximity is reported as an identifier for the locus, we emphasize that we are not implying that this gene is the causal gene for AD pathogenesis. We aim to highlight the most likely causal genes with more sophisticated functional interpretation analyses in later sections of this study.

available in the case-control meta-analysis and 100% of these were sign-concordant (*P*=0.0020). Such substantial overlap suggests that the AD-by-proxy phenotype captures a large part of the genetic effects on AD.

Given the high genetic overlap, in phase 3 we conducted a meta-analysis on the clinical AD case-control GWAS and the AD-by-proxy GWAS (**Figure 1**), comprising a total sample size of 455,258 (N_{eff}=450,734), including 71,880 (proxy) cases and 383,378 (proxy) controls. The LD score intercept was 1.0018 (SE=0.0109) and the λ_{1000} was 1.044, indicating again that most of the inflation in genetic signal (λ_{GC}=1.0833) could be explained by polygenicity (**Supplementary Figure 1b**). There were 2,357 GWS variants, which were represented by 94 lead SNPs, located in 29 distinct loci (**Table 1**, **Figure 2, Supplementary Figure 2**). These included 15 of the 18 loci detected in our case-control analyses, all of the 13 detected in the AD-by-proxy analyses, as well as 9 loci that were sub-threshold in both individual analyses but reached significance in the meta-analysis. A large proportion of the lead SNPs (60/94) were 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 LD-independent SNPs likely to represent a single association signal. The top lead SNP in every locus demonstrated concordant directions of effect in both the case-control and AD-by-proxy analyses for which they were both available (27 of 29 loci). Further, for 22 (out of 27 overlapping) loci, a robust GWAS association was observed independently in both the case-control and AD-by-proxy results, as defined by one or more SNPs in the locus having a *P*-value less than $5.3\text{x}10^{-4}$ (0.05/94 LD-independent signals) in both analyses. The effect size estimates appeared sometimes lower and sometimes higher for AD-by-proxy than for AD but were generally consistent in direction.

To confirm that the 29 significantly associated genomic loci are genuinely independent risk signals, we tested each locus for its association to AD while conditioning on significant top lead SNP(s) of other loci on the same chromosome. All, except one locus on chromosome 19, remained significant after the conditional analysis (**Supplementary Table 3**). The novel locus *AC074212.3* on chromosome 19 (locus #27) showed a modest reduction in signal after conditioning on the *APOE* top SNP (from P=4.64x10^{-8} to P=5.61x10^{-7}), implying that most but not all of the association is independent of *APOE.* In contrast, conditional analysis of the 6 loci with multiple lead SNPs showed that for 3 loci (*TREM2, PTK2B/CLU* and *APOE*), at least 1 significantly associated SNP remained within the locus after conditioning on the most significant lead SNP (**Supplementary Table 4**). This implies that for these loci, multiple causal signals might exist, putatively contributing to AD through distinct biological mechanisms. Specifically, for the *APOE* region (locus #26), 8 GWS signals remained after conditioning on successive sets of the strongest SNPs in the locus. For the other 3 loci, multiple lead SNPs in close proximity (250kb) were best captured by a single association signal, validating the decision to collapse these into single risk loci. Although this finding is important to recognize for future studies, detailed definition of multiple signals within the risk loci goes beyond the scope of the current study, and we therefore study the SNPs in our subsequent functional analysis within the 29 defined loci.

Of the 29 associated loci, 16 overlapped one 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 data. Re-analysis of these loci, while excluding the ADSP dataset, resulted in same strength association signals (**Supplementary Table 5**), implying that we have sufficiently adjusted for partial sample overlap between IGAP and ADSP in the Phase III meta-analysis. The lead SNPs in three loci (with nearest genes *HESX1*, *TREM2* and *CNTNAP2*) were only available in the UKB cohort (**Table 1**), but were of good quality (INFO>0.91, HWE *P*>.19, missingness<.003). These SNPs were all rare (MAF < .003) and thus could only be tested in a large cohort like UKB, meaning that they will require future confirmation in another similarly large sample.

Verifying our results against other[9,18] and more recent[12,16,19] genetic studies on AD, 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 (*ADAMTS4, HESX1, CLNK, CNTNAP2, ADAM10, APH1B, KAT8, ALPK2, AC074212.3*). Comparing our meta-analysis results of Phase I, Phase II and Phase III 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 1.6x10^{-5} to 0.0011), which was mostly caused by a lower association signal in the UKB dataset (**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* is also not supported (meta-analysis *P*=0.030; UKB analysis *P*=0.766; **Table 1**), implying invalidation of this locus, while the *CD33* locus (rs3865444 in **Table 1**) is significantly associated with AD (meta-analysis P=6.34 x 10^{-9}; UKB analysis P=4.97 x 10^{-5}), implying a genuine genetic association with AD risk.

Next, we aimed to find further support for the novel findings of the phase 3 meta-analysis by using an independent Icelandic cohort (deCODE[20,21]), including 6,593 AD cases and 174,289 controls (**Figure 1;** see **Methods**; **Supplementary Table 7**) to test replication of the lead SNP or a LD-proxy of the lead SNP (r^2>.9) in each locus. We like to note though that a formal, single locus level replication would require an independent dataset which is larger than the discovery dataset.

Given our large discovery sample size such a dataset is currently not available. This is a general issue in the current GWAS field where discovery sets are increasingly large. To have at least an indication of effects of GWS SNPS in an independent sample our replication effort is mainly intended to show whether there is sign concordance and enrichment of low P-values in the loci in general.

We were unable to test two loci as the lead SNPs (and SNPs in high LD), either were not present in the 28,075 genomes of 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*, *AC074212.3*) showed nominally significant association results (P<0.05) for the same SNP or a SNP in high LD (r^2 > 0.9) within the same locus (two-tailed binomial test P=1.9x10^{-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 (P=4.17x10^{-7}) that were available in deCODE (26 out of 27). As an additional method of testing for replication, we used genome-wide polygenic score prediction in two independent samples.[22] The current results explain 7.1% of the variance in clinical AD at a low best fitting P-threshold of 1.69x10^{-5} (P=1.80x10^{-10}) in 761 individuals with case-control diagnoses (see **Methods**). When excluding the *APOE*-locus (chr19: 45020859-45844508), the results explain 3.9% of the variance with a best fitting P-threshold of 3.5x10^{-5} (P=1.90x10^{-6}). We also predict AD status in a sample of 1459 pathologically confirmed cases and controls[23] with an R^2=0.41 and an area under the curve (AUC) of 0.827 (95% CI: 0.805-0.849, P=9.71x10^{-70}) using the best-fitting model of SNPs with a GWAS P<.50, as well as R^2=0.23

and AUC=0.733 (95% CI: 0.706-0.758, P=1.16x10^{-45}) using only *APOE* SNPs. This validation sample contains a small number (< 2%) 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%.[23]

Functional interpretation of genetic variants contributing to AD and AD-by-proxy

Next, we conducted a number of *in silico* follow-up analyses to interpret our findings in a biological context. Functional annotation of all GWS SNPs (n=2,357) in the associated loci showed that SNPs were mostly located in intronic/intergenic areas, yet in regions that were enriched for chromatin states 4 and 5, implying effects on active transcription (**Figure 3A, 3B and 3C; Supplementary Table 8**). 25 GWS SNPs were exonic non-synonymous (ExNS) (**Figure 3A; Supplementary Table 9**) with likely deleterious implications on gene function. Converging evidence of strong association (Z> |7|) and a high observed probability of a deleterious variant effect (CADD[24] 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 catalogue of variants in the associated genomic loci. We also applied a fine-mapping model[25] 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 out of 854 SNPs were selected in the *APOE* locus (#26), while 345 out of 434 SNPs were nominated in the *HLA-DRB1* locus (#7). Credible causal SNPs were not limited to known functional categories such as ExNS, indicating more complicated causal pathways that merit investigation with the set of variants prioritized by these statistical and functional annotations.

Partitioned heritability analysis,[26] excluding SNPs with extremely large effect sizes (i.e. *APOE* variants) showed enrichment for h^2_{SNP} for variants located in H3K27ac marks (Enrichment=3.18, P=9.63×10^{-5}), which are associated with activation of transcription, and in Super Enhancers (Enrichment=3.62, P=2.28×10^{-4}), which are genomic regions where multiple epigenetic marks of active transcription are clustered (**Figure 3D; Supplementary Table 10**). Heritability was also enriched in variants on chromosome 17 (Enrichment=3.61, P=1.63x10^{-4}) and we observed a trend of enrichment for heritability in common rather than rarer variants (**Supplementary Figure 3; 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 nonsynonymous 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 FUMA[27] (**Online Methods**). We used all SNPs with a P-value < 5x10^{-8} for gene-mapping. *Positional* gene-mapping aligned SNPs to 100 genes by their location within or immediately up/downstream (+/-10kb) of known gene boundaries, *eQTL (expression quantitative trait loci)* gene-mapping matched cis-eQTL SNPs to 171 genes whose expression levels they influence in one or more tissues, and c*hromatin interaction* mapping linked SNPs to 22 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 Figure 4; Supplementary Tables 13 and 14**). This resulted in 195 uniquely mapped genes, 81 of which were implicated by at least two mapping strategies and 17 by all 3 (**Figure 4E**). 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**). Of special interest is the locus on chromosome 8 (*CLU/PTK2B*). In the GWAS by Lambert et al.[4], this locus was defined as 2 distinct loci (*CLU* and *PTK2B*). Although our conditional analysis based on genetic data also specified this locus as having at least 2 independent association signals (**Supplementary Table 4)**, the chromatin interaction data in two immune-related tissues – the spleen and liver (**Supplementary Table 14**), suggests that the *PTK2B* and *CLU* loci might physically interact with their genomic regions (**Figure 3E**), therefore putatively affecting AD pathogenesis via the same biological mechanism. Future studies should thus consider the joint effects of how these two genes simultaneously impact AD risk. Chromosome 16 contains a locus implicated by long-range eQTL association (**Figure 3F**) clearly illustrating how the more distant genes *C16orf93, RNF40* and *ITGAX* can be affected by a genetic factor (rs59735493) in various body tissues (e.g. blood, skin), including a change in expression for *RNF40* observed in the dorsolateral prefrontal cortex (see next section). These observations emphasize the relevance of considering putative causal genes or regulatory elements not solely on the physical location but also on epigenetic influences. The identified susceptibility loci contained a substantially enriched proportion of eQTL

SNPs (80.7% versus 21.4% of genomic SNPs in the annotation databases, hypergeometric test $P < 1x10^{-324}$), indicating that such regulatory effects are prevalent in the genetic association signal. **Supplementary Figure 4** displays chromatin interaction patterns for all chromosomes containing significant GWAS loci.

In addition to the FUMA gene-mapping strategy that applies a general annotation approach including all body tissues, we explored the putative involvement of genes in AD pathogenesis through QTL annotation (expression (eQTL); methylation (mQTL) and histone (haQTL)) based on brain-specific public databases, including BRAINEAC[28], CommonMind Consortium Portal[29] and xQTL Serve[30]. For 16 of the 29 loci, a significant eQTL was observed for at least 1 of the 10 studied brain regions (**Supplementary Table 15**). Focusing on brain regions that are typically degenerated in brains from patients with AD, we observed a change in expression for *GATS* in the temporal cortex and for *ZNF789* and *PILRB* in the hippocampus for locus 10. For the *APOE*-locus, a change in expression for *NKPD1* was reported in the hippocampus. For the *HLA*-locus, *HLA-DQA2* expression was significantly influenced by the LD-block (represented by 108 variants) in the hippocampus. Using xQTL server, we identified significant mQTLs in the dorsolateral prefrontal cortex for 10 loci (**Supplementary Table 16**). Again, *STAG3* and *PILRB* are implied to be affected by genomic locus 10, though this time to changes in methylation levels in close proximity to these genes. None of the 29 significant loci overlapped with annotated haQTLs.

For the above reported eQTL associations in both the general and the brain-specific analysis, the eQTL SNPs themselves are significantly associated to AD in Phase III. Furthermore, we confirmed that these reported eQTLs appear to co-localize in a moderate to high manner with the lead SNP of the locus of interest, as the association signals for eQTL SNPs were substantially attenuated when controlling for non-eQTL lead SNPs in the same loci (**Supplementary Table 17**).

Although these gene-mapping strategies imply multiple putative causal genes per GWAS locus, several of these genes in the novel loci (and significantly replicated by the deCODE cohort) are of particular interest, as the genes have functional or previous genetic association to 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 their role in AD pathogenesis.[31] 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*,[32] However this is the first time that this gene is implicated as a common risk factor for AD, and 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 nonsynonymous variant in exon 1 of *APH1B,* which encodes for a protein subunit of the *γ*-secretase complex cleaving *APP.*[33] 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 levels of *KAT8* in multiple brain regions (**Supplementary Tables 13 and 16**) including hippocampus. The chromatin modifier *KAT8* is regulated by *KANSL1,* a gene associated to AD in absence of APOE ε4. A study on Parkinson's disease (PD) reported *KAT8* as potential causal gene based on GWAS and differential gene expression results, implying a putative shared role in neurodegeneration of *KAT8* in AD and PD.[34] 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 MAGMA.[35] This method annotates SNPs to known protein-coding genes to estimate aggregate associations based on all SNPs in a gene. It differs from the gene-mapping strategies in 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 Figure 5; Supplementary Table 18**), of which 78 were also mapped by FUMA (**Figure 4E**). In total, 16 genes were implicated by all four 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 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 6,994 biological-pathway-based gene-sets, 53 tissue expression-based gene-sets and 39 brain single-cell expression based gene-sets (24 derived from mouse data and 15 derived from human data). We found four Gene Ontology[19] gene-sets that were significantly associated with AD risk: *Protein lipid complex* (P=3.93×10^{-10}), *Regulation of amyloid precursor protein catabolic process* (P=8.16×10^{-09}), *High density lipoprotein particle* (P=7.81x10^{-8}), and *Protein lipid complex assembly* (P=7.96×10^{-7}) **(Figure 4A; Supplementary Tables 20 and 21)**. Conditional analysis on the *APOE* locus showed associations with AD for these four gene-sets independent of the effect of *APOE*, as they remained significantly associated (P<0.0125), though somewhat decreased in strength, suggesting that *APOE* is contributing a substantial part to the association signal but does not completely drive the effect. There was overlap between genes included in the four gene-sets, and conditioning on each significant gene-set association showed that three gene-sets were associated with AD independently of each other (**Supplementary Tables 20 and 21**). All 25 genes of the *High density lipoprotein particle* pathway are also part of the *Protein lipid complex* (conditional analysis *P*=0.18)*,* and these pathways are therefore not interpretable as independent associations.

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 (i.e. 6,994 GO categories, 54 tissues and 39 cell types). However, we did observe suggestive associations when correcting only for the number of tests within all tissue types or cell-types. This was the case for gene expression across immune-related tissues (**Figure 4C; Supplementary Table 22**), particularly whole blood (*P*=5.61×10^{-6}), spleen (*P*=$1.50\text{x}10^{-5}$) and lung (*P*=$4.67\text{x}10^{-4}$), 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 dataset (*P*=$1.96\text{x}10^{-3}$), though not surviving the stringent Bonferroni correction (**Figure 4B; Supplementary Table 23**). However, we observed a similar association signal for microglia in a second independent single-cell expression dataset in humans (*P*=$2.56\text{x}10^{-3}$) (**Supplementary Figure 6; 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

For a more comprehensive understanding of the genetic background of AD, we next tested whether AD is likely to share genetic factors with other phenotypes. To determine the existence of putative pleiotropic effects, we annotated the GWS SNPs with information from the NCBI GWAS catalog[36] using FUMA. Seven loci (locus numbers 1, 6 ,9, 15, 16, 26, 28) were associated with non-AD related phenotypes (**Supplementary Table 25**), including the *HLA* (locus 6) and *APOE*-locus (locus 26). Besides these two loci that have been extensively linked to a variety of diseases due to their large LD-blocks, we observed previously reported associations fir immune-related, blood-related and lipid-related phenotypes. One novel locus (locus 1) has been earlier associated with monocyte percentage of white cells.

We furthermore conducted bivariate LD score[14] regression to test for genetic correlations between AD and 40 other traits for which large GWAS summary statistics were available. We observed significant negative genetic correlations with adult cognitive ability (r_g=-0.22, P=7.28x10^{-5}), age of first birth (r_g=−0.33, P=1.22×10^{-4}) and educational attainment (r_g=−0.25, P=5.01×10^{-4}) (**Figure 4D; Supplementary Table 26**).

We then used Generalised Summary-statistic-based Mendelian Randomisation[37] (GSMR; see **Methods**) 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 (i.e. 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 (OR) for AD as a function of every SD increase in the exposure phenotype. We observed a protective effect of cognitive ability (OR=0.89, 95% CI: 0.85-0.92, P=5.07x10^{-9}), educational attainment (OR=0.88, 95%CI: 0.81-0.94, P=3.94×10^{-4}), and height (OR=0.96, 95%CI: 0.94-0.97, P=1.84x10^{-8}) on risk for AD (**Supplementary Table 27; Supplementary Figure 7)**. No substantial evidence of pleiotropy was observed between AD and these phenotypes, with <1% of overlapping SNPs being filtered as outliers (**Supplementary Figure 7**).

Discussion

By using an unconventional approach of including a proxy phenotype for AD to increase sample size, we have identified 9 novel loci and gained novel biological knowledge on AD aetiology. We were able to test 7 of the 9 novel loci for replication, of which 4 loci showed clear replication, 1 locus showed marginal replication and 2 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 loci 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,[19,38] we emphasise 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, 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.[39] 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 prior 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.[40]

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 prior existing cognitive processing strategies.[41] 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.[42] 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[43,44], a measure associated to decreased risk of AD.[45] This indirect association is furthermore supported by the observed increase in cognitive reserve for taller individuals.[46] Alternatively, genetic variants influencing height might also affect biological mechanisms involved in AD aetiology, such as *IGF1* that codes for the insulin-like growth factor and is associated with cerebral amyloid.[47]

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

URLs:

http://ukbiobank.ac.uk

https://www.ncbi.nlm.nih.gov/gap

http://fuma.ctglab.nl

http://ctg.cncr.nl/software/magma

http://genome.sph.umich.edu/wiki/METAL_Program

https://github.com/bulik/ldsc

http://ldsc.broadinstitute.org/

https://data.broadinstitute.org/alkesgroup/LDSCORE/

http://www.genecards.org

http://www.med.unc.edu/pgc/results-and-downloads

http://software.broadinstitute.org/gsea/msigdb/collections.jsp

https://www.ebi.ac.uk/gwas/

https://github.com/ivankosmos/RegionAnnotator

http://cnsgenomics.com/software/gsmr/

https://github.com/hailianghuang/FM-summary

Acknowledgments: This work was funded by The Netherlands Organization for Scientific Research (NWO VICI 453-14-005) and the Sophia Foundation for Scientific Research (grant nr: S14-27). 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 (#251134, #248778, #223273, #213837, #225989), KG Jebsen Stiftelsen, The Norwegian Health Association, European Community's JPND Program, ApGeM RCN #237250, and the European Community's grant # PIAPP-GA-2011-286213 PsychDPC. This research has been conducted using the UK Biobank resource under application number 16406 and the public ADSP dataset, obtained through the Database of Genotypes and Phenotypes (dbGaP) under accession number phs000572 (see sections below).

Genotyping for the Swedish Twin Studies of Aging was supported by NIH/NIA grant R01 AG037985. Genotyping in TwinGene was supported by NIH/NIDDK U01 DK066134. WvdF is recipient of Joint Programming for Neurodegenerative Diseases (JPND) grants PERADES (ANR-13-JPRF-0001) and EADB (733051061). AddNeuroMed consortium was led by Simon Lovestone, Bruno Vellas, Patrizia Mecocci, Magda Tsolaki, Iwona Kłoszewska, Hilkka Soininen. This work was supported by InnoMed (Innovative Medicines in Europe), an integrated project funded by the European Union of the Sixth Framework program priority (FP6-2004- LIFESCIHEALTH-5). JB was supported by a grant from the Swiss National Science Foundation. JHL was supported by the Swedish Research Council (Vetenskapsrådet, award 2014-3863), the Wellcome Trust (108726/Z/15/Z), and the Swedish Brain Foundation (Hjärnfonden). NS was supported by the Wellcome Trust (108726/Z/15/Z). RD was supported by National Institute for Health Research University College London Hospital’s Biomedical Research Centre, Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the Chief Scientist Office, the Economic and Social Research Council, the Engineering and Physical Sciences Research Council, the National Institute for Social Care and Health Research, and the Wellcome Trust (grant number MR/K006584/1), Innovative Medicines Initiative Joint Undertaking under EMIF grant agreement number 115372, resources of which are composed of financial contribution from the European Union's Seventh Framework Program (FP7/2007-2013) and EFPIA companies' in kind contribution. SJK was supported by an MRC Career Development Award in Biostatistics (MR/L011859/1).

We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The i–Select chips was funded by the French National Foundation on Alzheimer's disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant n° 503480), Alzheimer's Research UK (Grant n° 503176), the Wellcome Trust (Grant n° 082604/2/07/Z) and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant n° 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01–AG–12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC–10–196728. This paper represents independent research funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation

Trust and King’s College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

The Alzheimer’s Disease Sequencing Project (ADSP) is comprised of two Alzheimer’s Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer’s Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. Seshadri and Fornage. Data generation and harmonization in the Follow-up Phases is supported by U54AG052427 (to Drs. Schellenberg and Wang). The ADGC cohorts include: Adult Changes in Thought (ACT), the Alzheimer’s Disease Centers (ADC), the Chicago Health and Aging Project (CHAP), the Memory and Aging Project (MAP), Mayo Clinic (MAYO), Mayo Parkinson’s Disease controls, University of Miami, the Multi-Institutional Research in Alzheimer’s Genetic Epidemiology Study (MIRAGE), the National Cell Repository for Alzheimer’s Disease (NCRAD), the National Institute on Aging Late Onset Alzheimer's Disease Family Study (NIA-LOAD), the Religious Orders Study (ROS), the Texas Alzheimer’s Research and Care Consortium (TARC), Vanderbilt University/Case Western Reserve

University (VAN/CWRU), the Washington Heights-Inwood Columbia Aging Project (WHICAP) and the Washington University Sequencing Project (WUSP), the Columbia University Hispanic-Estudio Familiar de Influencia Genetica de Alzheimer (EFIGA), the University of Toronto (UT), and Genetic Differences (GD). The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193. The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904),the EU Joint Programme - Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435. ARIC research is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01-

HL70825 from the NHLBI. CHS research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS (R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly

(014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their participants, faculty and staff. The content of these manuscripts is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services. The three LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), and the Washington University Genome Institute (U54HG003079). Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Cell Repository for Alzheimer's Disease (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Study Investigators, the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC, U01AG016976) and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA, and at the Database for Genotypes and Phenotypes (dbGaP) funded by NIH. This research was supported in part by the Intramural Research Program of the National Institutes of health, National Library of Medicine. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or nongovernmental organizations.

We thank the numerous participants, researchers, and staff from many studies who collected and contributed to the data. Summary statistics will be made available for download upon publication from http://ctglab.vu.nl.

Author Contributions: I.E.J. and J.E.S. performed the analyses. D.P. and O.E.A. conceived the idea of the study. D.P. and S.R. supervised analyses. Sv.St. performed QC 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.

Author Information: Patrick F Sullivan reports the following potentially competing financial interests: Lundbeck (advisory committee), Pfizer (Scientific Advisory Board member), and Roche (grant recipient, speaker reimbursement). Jens Hjerling-Leffler: Cartana (Scientific Advisor) and Roche (grant recipient). Ole A Andreassen: (Lundbeck) speaker's honorarium. Stacy Steinberg, Hreinn Stefansson and Kari Stefansson are employees of deCODE Genetics/Amgen. John Hardy is a cograntee of Cytox from Innovate UK (U.K. Department of Business). All other authors declare no financial interests or potential conflicts of interest.

Correspondence and requests for materials should be addressed to d.posthuma@vu.nl.

Tables

Table 1. Summary statistics for the meta-analysis of case-control status, by proxy phenotype and both.

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	1.57E-04	rs4575098	6.88E-08	rs4575098	161155392	A	G	0.240	6.36	**2.05E-10**	?+++
2	1	*CR1*	rs6656401	**1.39E-17**	rs679515	**8.85E-10**	rs2093760	207786828	A	G	0.205	8.82	**1.10E-18**	++++
3	2	*BIN1*	rs4663105	**3.58E-29**	rs4663105	**5.46E-26**	rs4663105	127891427	C	A	0.415	13.94	**3.38E-44**	?+++
4	2	*INPPD5*	rs10933431	1.67E-06	rs10933431	2.51E-06	rs10933431	233981912	G	C	0.235	-6.13	**8.92E-10**	?---
5	3	***HESX1***	NA		rs184384746	**1.24E-08**	rs184384746	57226150	T	C	0.002	5.69	**1.24E-08**	???+
6	4	***CLNK***	rs6448453	0.024	rs6448451	**1.19E-08**	rs6448453	11026028	A	G	0.252	6.00	**1.93E-09**	?+-+
--	4	*HS3ST1*	rs7657553	**2.16E-08**	rs7657553	0.79	rs7657553	11723235	A	G	0.291	1.95	0.051	?++-
7	6	*HLA-DRB1*	rs9269853	**2.66E-08**	rs6931277	1.78E-07	rs6931277	32583357	T	A	0.153	-6.49	**8.41E-11**	?---
8	6	*TREM2*	NA		rs187370608	**1.45E-16**	rs187370608	40942196	A	G	0.002	8.26	**1.45E-16**	???+
9	6	*CD2AP*	rs9381563	**5.35E-09**	rs9381563	8.10E-06	rs9381563	47432637	C	T	0.355	6.33	**2.52E-10**	?+++
10	7	*ZCWPW1*	rs1859788	**6.05E-09**	rs7384878	**2.38E-10**	rs1859788	99971834	A	G	0.310	-7.93	**2.22E-15**	----
11	7	*EPHA1*	rs11763230	**2.58E-11**	rs7810606	1.01E-06	rs7810606	143108158	T	C	0.500	-6.62	**3.59E-11**	?---
12	7	***CNTNAP2***	NA		rs114360492	**2.10E-09**	rs114360492	145950029	T	C	0.000	5.99	**2.10E-09**	???+
13	8	*CLU/PTK2B*	rs4236673	**6.36E-20**	rs1532278	**7.45E-09**	rs4236673	27464929	A	G	0.391	-8.98	**2.61E-19**	----
14	10	*ECHDC3*	rs11257242	**2.38E-08**	rs11257238	5.84E-05	rs11257238	11717397	C	T	0.375	5.69	**1.26E-08**	?+++
15	11	*MS4A6A*	rs7935829	**8.21E-13**	rs1582763	**4.72E-09**	rs2081545	59958380	A	C	0.381	-7.97	**1.55E-15**	----
16	11	*PICALM*	rs10792832	**1.12E-17**	rs3844143	**5.31E-11**	rs867611	85776544	G	A	0.314	-8.75	**2.19E-18**	?---
17	11	*SORL1*	rs11218343	**5.57E-11**	rs11218343	2.81E-06	rs11218343	121435587	C	T	0.040	-6.79	**1.09E-11**	?---
18	14	*SLC24A4*	rs12590654	**1.98E-08**	rs12590654	3.70E-06	rs12590654	92938855	A	G	0.344	-6.39	**1.65E-10**	?---
19	15	***ADAM10***	rs442495	3.09E-04	rs442495	2.65E-07	rs442495	59022615	C	T	0.320	-6.07	**1.31E-09**	?---
20	15	***APH1B***	rs117618017	0.022	rs117618017	2.64E-07	rs117618017	63569902	T	C	0.132	5.52	**3.35E-08**	++++
21	16	***KAT8***	rs59735493	8.25E-04	rs59735493	3.72E-06	rs59735493	31133100	A	G	0.300	-5.49	**3.98E-08**	?---
22	17	*SCIMP*	rs113260531	3.21E-06	rs9916042	**4.73E-08**	rs113260531	5138980	A	G	0.120	6.12	**9.16E-10**	?+++
23	17	*ABI3*	rs28394864	7.29E-05	rs28394864	6.80E-06	rs28394864	47450775	A	G	0.473	5.62	**1.87E-08**	?+++
--	17	*BZRAP1-AS1*	rs2632516	**1.42E-09**	rs2632516	0.005	rs2632516	56409089	C	G	0.455	-4.90	9.66E-07	?---

--	18	*SUZ12P1*	rs8093731	**4.63E-08**	rs8093731	0.766	rs8093731	29088958	T	C	0.010	-2.17	0.03	?-?-
24	18	***ALPK2***	rs76726049	0.039	rs76726049	1.83E-07	rs76726049	56189459	C	T	0.014	5.52	**3.30E-08**	?+++
25	19	*ABCA7*	rs4147929	**8.64E-09**	rs3752241	**2.87E-08**	rs111278892	1039323	G	C	0.161	6.50	**7.93E-11**	?+++
26	19	*APOE*	rs41289512	**2.70E-194**	rs75627662	**9.51E-296**	rs41289512	45351516	G	C	0.039	35.50	**5.79E-276**	?+++
27	19	***AC074212.3***	rs76320948	1.54E-05	rs76320948	1.80E-05	rs76320948	46241841	T	C	0.046	5.46	**4.64E-08**	?+?+
28	19	*CD33*	rs3865444	**4.25E-08**	rs3865444	4.97E-05	rs3865444	51727962	A	C	0.320	-5.81	**6.34E-09**	?---
29	20	*CASS4*	rs6014724	8.72E-08	rs6014724	6.32E-06	rs6014724	54998544	G	A	0.089	-6.18	**6.56E-10**	?---

Note: Independent lead SNPs are defined by r2 < .1; distinct genomic loci are >250kb apart. The locus column indicates the loci number based on Phase III (-- indicates that this locus is non-significant). The strongest lead SNP from each locus in the Phase III meta-analysis is presented here. 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 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, UKB note that the first cohort is often missing as this concerns exome sequencing data. Corrected P value for significance = 5E-08 (marked as bold and underlined values). Note that the lead SNP can differ between the distinct analyses, while it tags the same locus.

Figure 1. Overview of analyses 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 dataset.

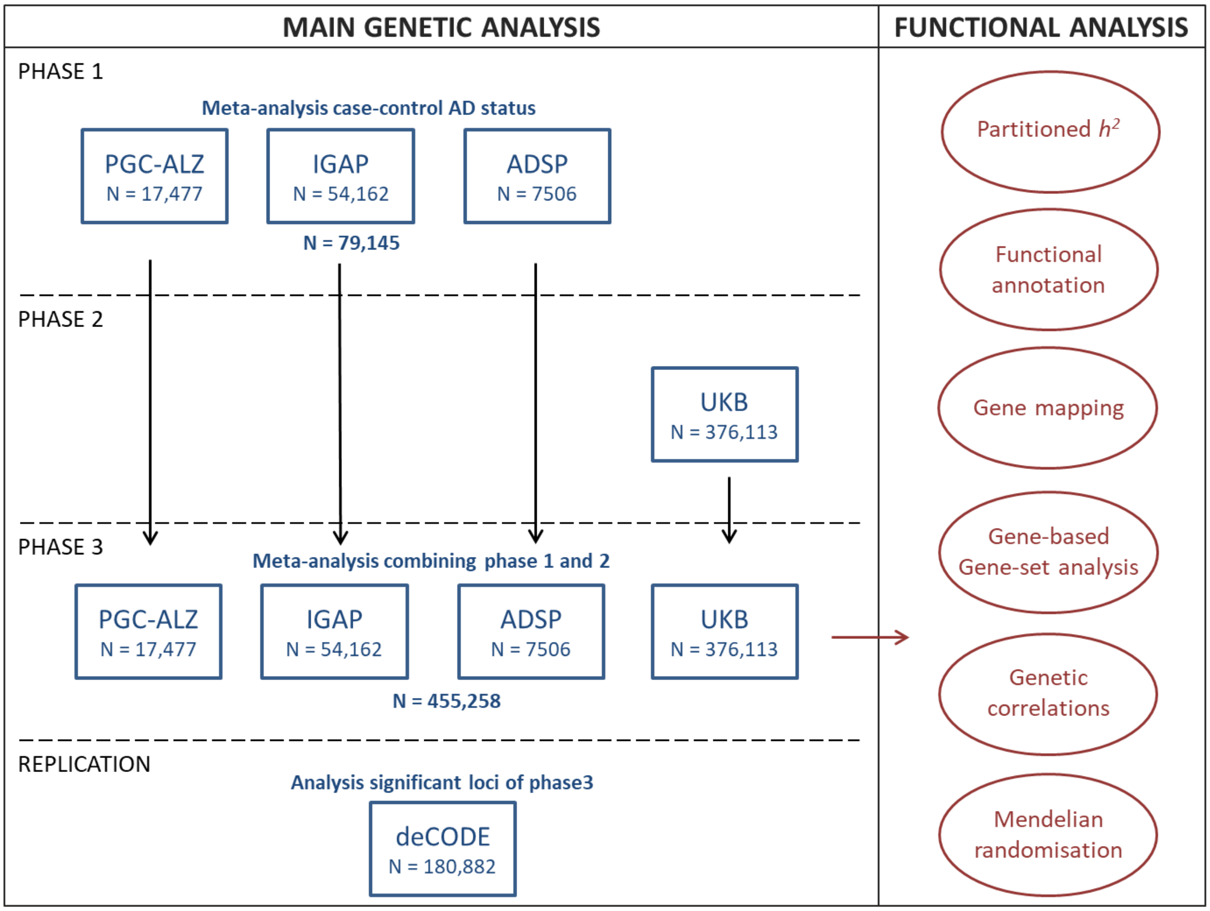
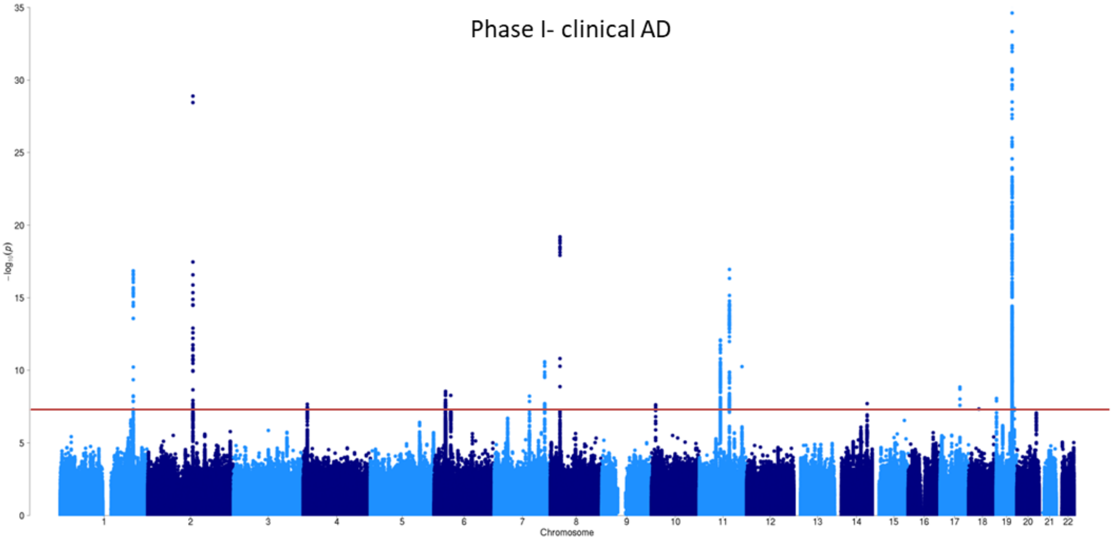
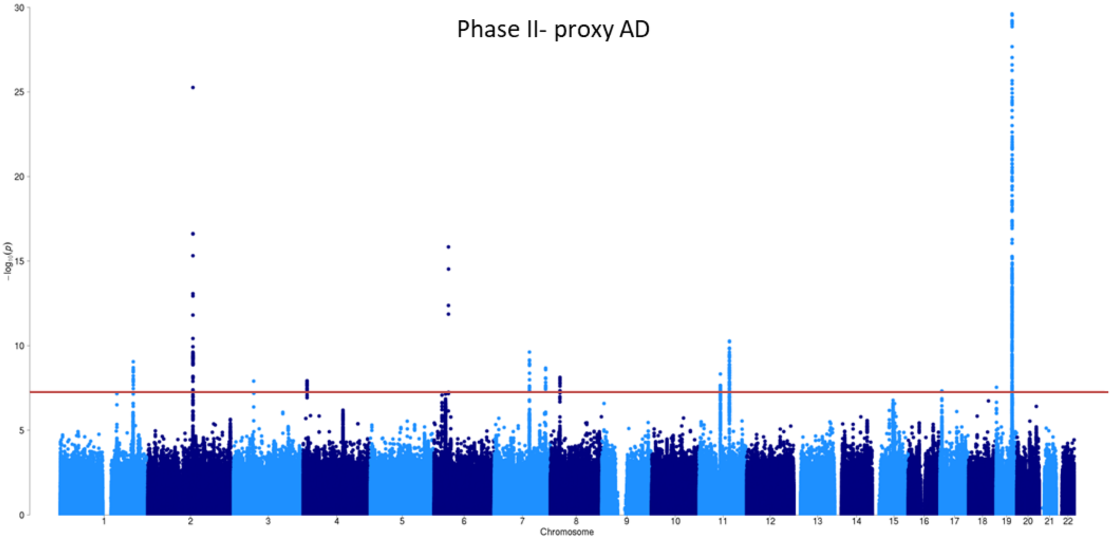


Figure 2. GWAS results for AD risk (N=455,258). Manhattan plot displays all associations per variant ordered according to their genomic position on the x-axis and showing the strength of the association with the –log10 transformed P-values on the y-axis. The y-axis is limited to enable visualization of non-*APOE* loci. For the Phase III meta-analysis, the original –log10 P-value for the APOE locus is 276.


Phase I- clinical AD
35
30
25
20
15
10
5
0
-log10(p)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome


Phase II- proxy AD
30
25
20
15
10
5
0
-log10(p)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

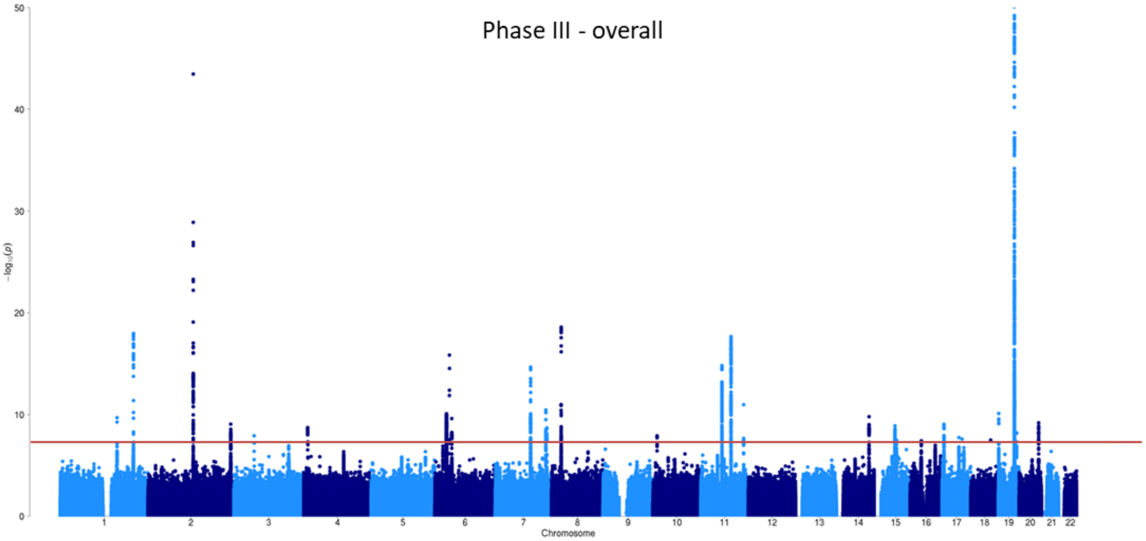

Phase III - overall
50
40
30
20
10
0
-log10(p)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

Figure 3. Functional annotation of association results. a) Functional effects of variants in genomic risk loci of the meta-analysis (the colours 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; **c)** Distribution of minimum chromatin state across 127 tissue and cell types for variants in genomic risk loci, with lower states indicating higher accessibility and states 1-7 referring to open chromatin states; **d)** Heritability enrichment of 28 functional variant annotations calculated with stratified LD score regression. UTR=untranslated region; 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 blue areas indicate genomic risk loci, green indicates eQTL associations and orange indicates chromatin interactions. Genes mapped by both eQTL and chromatin interactions are red. The outer layer shows a Manhattan plot containing the negative log10-transformed P-value of each SNP in the GWAS meta-analysis of AD. Full circos plots of all autosomal chromosomes are provided in Supplementary Figures 4.

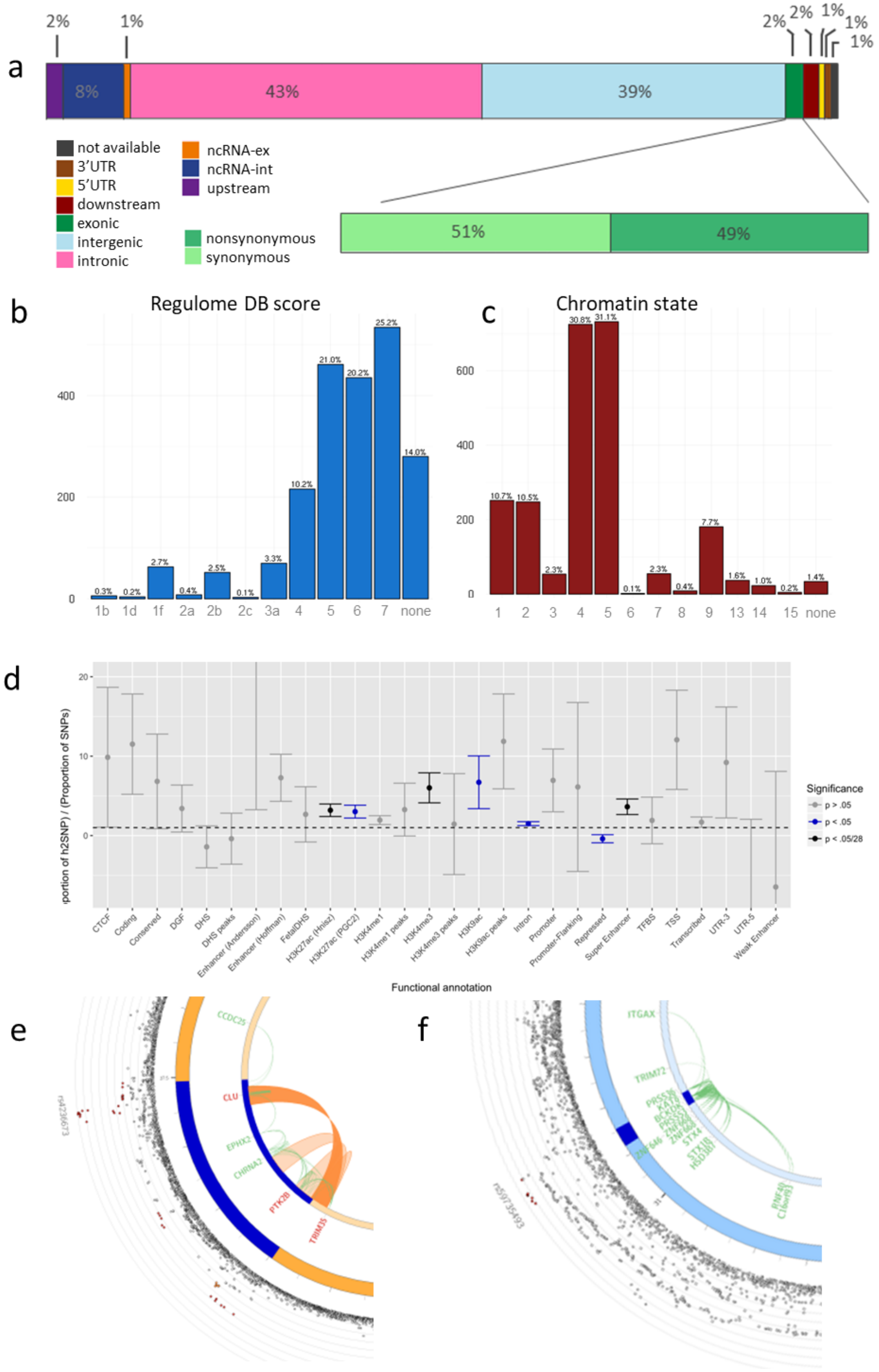
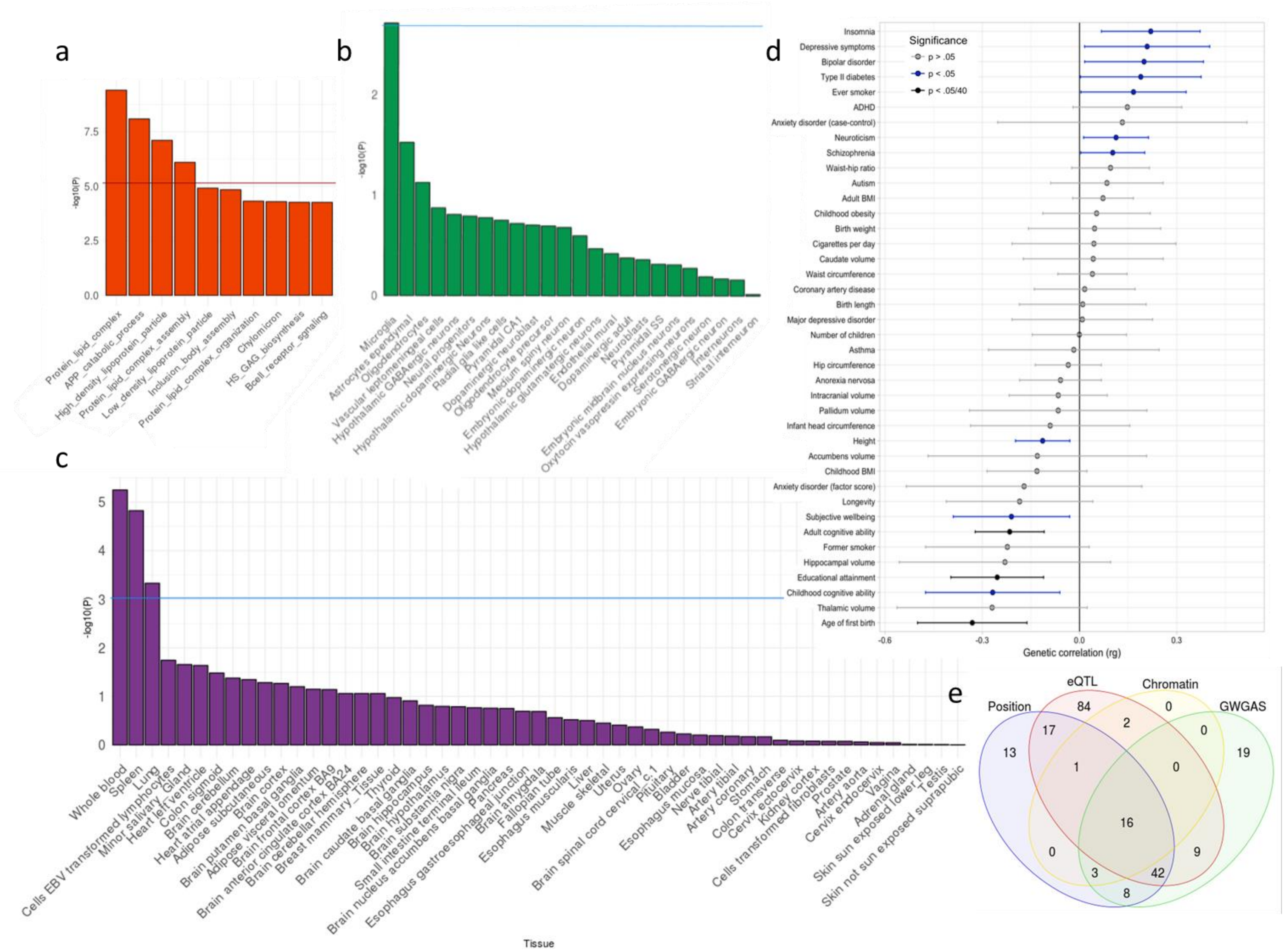

a
2%
1%
8%
43%
39%
2%
2%
1%
1%
1%
not available
3'UTR
5'UTR
downstream
exonic
intergenic
intronic
ncRNA-ex
ncRNA-int
upstream
nonsynonymous
synonymous
51%
49%
b
Regulome DB score
0
200
400
0.3%
0.2%
2.7%
0.4%
2.5%
0.1%
3.3%
10.2%
21.0%
20.2%
25.2%
14.0%
1b
1d
1f
2a
2b
2c
3a
4
5
6
7
none
c
Chromatin state
0
200
400
600
10.7%
10.5%
2.3%
30.8%
31.1%
0.1%
2.3%
0.4%
7.7%
1.6%
1.0%
0.2%
1.4%
1
2
3
4
5
6
7
8
9
13
14
15
none
d
portion of h2SNP) / (Proportion of SNPs)
0
10
20
Significance
p > .05
p < .05
p < .05/28
CTCF
Coding
Conserved
DGF
DHS
DHS peaks
Enhancer (Andersson)
Enhancer (Hoffman)
FetalDHS
H3K27ac (Hnisz)
H3K27ac (PGC2)
H3K4me1
H3K4me1 peaks
H3K4me3
H3K4me3 peaks
H3K9ac
H3K9ac peaks
Intron
Promoter
Promoter-Flanking
Repressed
Super Enhancer
TFBS
TSS
Transcribed
UTR-3
UTR-5
Weak Enhancer
Functional annotation
e
rs4236673
CCDC25
CLU
EPHX2
CHRNA2
PTK2B
TRIM35
f
rs59735493
ITGAX
TRIM72
PRSS36
KAT8
BCKDK
PRSS53
ZNF668
ZNF646
ZNF646
STX4
STX1B
HSD3B7
RNF40
C16orf93
31

Figure 4. Functional implications based on gene-set analysis, genetic correlations and functional annotations. 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 (i.e. gene-ontology, tissue expression, single cell expression). (D) Genetic correlations between AD and other heritable traits. (E) Venn diagram showing the number of genes mapped by four distinct strategies.


a
-log10(P)
0.0
2.5
5.0
7.5
Protein_lipid_complex
APP_catabolic_process
High_density_lipoprotein_particle
Protein_lipid_complex_assembly
Low_density_lipoprotein_particle
Inclusion_body_assembly
Protein_lipid_complex_organization
Chylomicron
HS_GAG_biosynthesis
Bcell_receptor_signaling
b
-log10(P)
0
1
2
Microglia
Astrocytes ependymal
Oligodendrocytes
Vascular leptomeningeal cells
Hypothalamic GABAergic neurons
Neural progenitors
Hypothalamic dopaminergic Neurons
Radial glia like cells
Pyramidal CA1
Dopaminergic neuroblast
Oligodendrocyte precursor
Medium spiny neuron
Embryonic dopaminergic neuron
Hypothalamic glutamatergic neurons
Endothelial mural
Dopaminergic adult
Neuroblasts
Pyramidal SS
Embryonic midbrain nucleus neurons
Oxytocin vasopressin expressing neurons
Serotonergic neuron
Embryonic GABAergic neuron
Interneurons
Striatal interneuron
c
-log10(P)
0
1
2
3
4
5
Whole blood
Spleen
Lung
Cells EBV transformed lymphocytes
Minor salivary Gland
Heart left ventricle
Colon sigmoid
Brain cerebellum
Heart atrial appendage
Adipose subcutaneous
Brain cortex
Brain putamen basal ganglia
Adipose visceral omentum
Brain frontal cortex BA9
Brain anterior cingulate cortex BA24
Brain cerebellar hemisphere
Breast mammary_Tissue
Thyroid
Brain caudate basal ganglia
Brain hippocampus
Brain hypothalamus
Small intestine terminal ileum
Brain substantia nigra
Brain nucleus accumbens basal ganglia
Pancreas
Esophagus gastroesophageal junction
Brain amygdala
Fallopian tube
Esophagus muscularis
Liver
Muscle skeletal
Uterus
Brain spinal cord cervical_c.1
Ovary
Pituitary
Bladder
Esophagus mucosa
Nerve tibial
Artery tibial
Artery coronary
Stomach
Colon transverse
Cervix ectocervix
Cells transformed fibroblasts
Kidney cortex
Prostate
Artery aorta
Cervix endocervix
Vagina
Adrenal gland
Skin sun exposed lower leg
Testis
Skin not sun exposed suprapubic
Tissue
d
Significance
p > .05
p < .05
p < .05/40
Insomnia
Depressive symptoms
Bipolar disorder
Type II diabetes
Ever smoker
ADHD
Anxiety disorder (case-control)
Neuroticism
Schizophrenia
Waist-hip ratio
Autism
Adult BMI
Childhood obesity
Birth weight
Cigarettes per day
Caudate volume
Waist circumference
Coronary artery disease
Birth length
Major depressive disorder
Number of children
Asthma
Hip circumference
Anorexia nervosa
Intracranial volume
Pallidum volume
Infant head circumference
Height
Accumbens volume
Childhood BMI
Anxiety disorder (factor score)
Longevity
Subjective wellbeing
Adult cognitive ability
Former smoker
Hippocampal volume
Educational attainment
Childhood cognitive ability
Thalamic volume
Age of first birth
-0.6
-0.3
0.0
0.3
Genetic correlation (rg)
e
Position
eQTL
Chromatin
GWGAS
13
84
0
19
17
2
0
1
0
16
0
9
3
42
8

Online methods

1.1 Study Cohorts

1.1.1 PGC-ALZ cohorts

Three non-public datasets (the Norwegian DemGene network, The Swedish Twin Studies of Aging, and TwinGene) were meta-analyzed as part of the Alzheimer workgroup initiative of the Psychiatric Genomic Consortium (PGC-ALZ).

We collected genotype data from the Norwegian DemGene Network consisting of 2,224 cases and 1,855 healthy controls. The DemGene Study is a Norwegian network of clinical sites collecting cases from Memory Clinics based on standardised examination of cognitive, functional and behavioural measures and data on progression of most patients. We diagnosed 2,224 cases of AD from 7 studies: the Norwegian Register of persons with Cognitive Symptoms (NorCog), the Progression of Alzheimer's Disease and Resource use (PADR), the Dementia Study of Western Norway (DemVest), the AHUS study, the Dementia Study in Rural Northern Norway (NordNorge), the HUNT Dementia Stud, the Nursing Home study, and the TrønderBrain study. These cases were diagnosed according to the recommendations from the National Institute on Aging–Alzheimer's Association (NIA/AA) (AHUS), the NINCDS-ADRDA criteria (DemVest and TrønderBrain) or the ICD-10 research criteria (NorCog, PADR, NordNorge and HUNT). The controls from Norway were obtained through the AHUS, NordNorge, HUNT and TrønderBrain studies. The controls were screened with standardized interview and cognitive tests. Genotypes of the 4,079 individuals from the DemGene Study were obtained with Human Omni Express-24 v.1.1 (Illumina Inc., San Diego, CA, USA) at deCODE Genetics (Reykjavik, Iceland). To increase the statistical power of our association analysis, the controls were combined with additional 5786 population controls from Norwegian blood donor samples (Oslo University Hospital, Ullevål Hospital, Oslo) and controls from Thematically Organized Psychosis (TOP) Research Study (between 25-65 years). Control subjects of the TOP Research Study were of Caucasian origin without history of moderate/severe head injury, neurological disorder, mental retardation and were excluded if they or any of their close relatives had a lifetime history of a severe psychiatric disorder, a history of medical problems thought to interfere with brain function or significant illicit drug use.

The Swedish Twin Studies of Aging (STSA) (n cases = 398, n controls = 1079) includes three sub-studies of aging within the Swedish Twin Registry[48]: The Swedish Adoption/Twin Study of Aging (SATSA)[49], Aging in Women and MEN (GENDER)[50], and The Study of Dementia in Swedish Twins (HARMONY)[51]. Informed consent was obtained from all participants and the studies were approved by the Regional Ethics Board in Stockholm and the Institutional Review Board at the University of Southern California. DNA was extracted from blood samples and genotyped using Illumina Infinium PsychArray. Alzheimer’s disease patients were diagnosed as part of the studies according to the NINCDS/ADRDA criteria[52]. In addition, information on disease after last study participation was retrieved from three population-based health care registers: The National Patient Register, the Causes of Death Register, and the Prescribed Drug Register.

TwinGene[48] is a population-based study of older twins drawn from the Swedish Twin Registry. Written informed consent was obtained from all participants and the study was approved by the Regional Ethics Board in Stockholm. DNA was extracted from blood samples and genotyped using Illumina Human OmniExpress for 1,791 individuals. Information about

Alzheimer’s disease (n cases = 343, n controls = 9070) was extracted from the National Patient Register, the Causes of Death Register, and the Prescribed Drug Register, all of which are population-based health care registers with nationwide coverage.

1.1.2 IGAP

Publically available (http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php) genome-wide association analysis results of the International Genomics of Alzheimer's Project (IGAP)[4] were included as one of the four cohorts that were meta-analysed in our effort. IGAP is a large two-stage study based upon genome-wide association studies (GWAS) on individuals of European ancestry. We focused on the results of stage 1, for which IGAP used genotyped and imputed data of 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyse four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's disease Initiative – EADI, the Alzheimer Disease Genetics Consortium – ADGC, the Cohorts for Heart and Aging Research in Genomic Epidemiology consortium – CHARGE, the Genetic and Environmental Risk in AD consortium – GERAD). As the purpose of stage 2 (11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls) was replication of the significantly associated loci of stage 1, we limited the inclusion of the summary statistics for our own analyses to stage 1. Written informed consent was obtained from study participants or, for those with substantial cognitive impairment, from a caregiver, legal guardian or other proxy, and the study protocols for all populations were reviewed and approved by the appropriate institutional review boards.

1.1.3 ADSP

The Alzheimer's Disease Sequencing Project (ADSP) collaboration has the aim to identify novel genetic factors that contribute to AD risk by studying genetic sequencing data. ADSP has made their sequencing data available through the Genotypes and Phenotyps database (dbGaP) under the study accession: phs000572.v7.p (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000572.v1 .p1). We have obtained access to 10,907 individuals (5,771 cases, 5,136 controls) with whole-exome sequencing data to include as the second cohort within our meta-analysis. A substantial proportion of the ADSP individuals were previously also included in IGAP. We applied two strategies to prevent inflated meta-analysis results due to sample overlap: (1) exclusion of ADSP individuals that were duplicates based on genotype data comparison of individual level genetic data between IGAP and ADSP, (2) perform meta-analysis while correcting for cross-study LD score regression intercept (see section 1.4.). To accomplish the first approach we obtained access for all IGAP datasets for which individual level genotype data was available through dbGaP (phs000160.v1.p1 - https:// www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id= phs000160.v1.p1; phs000219.v1.p1 - https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000219.v1.p1; phs000372.v1.p1 - https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000372.v1 .p1; phs000168.v2.p2 - https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id= phs000168.v2.p2; phs000234.v1.p1 - https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi? study_id=phs000234.v1.p1) or NIAGADS (NG00026 -

https://www.niagads.org/datasets/ng00026; NG00028 - https://www.niagads.org/datasets/ng00028; NG00029 - https://www.niagads.org/ datasets/ng00029; NG00031 - https://www.niagads.org/datasets/ng00030 ; NG00031 - https://www.niagads.org/datasets/ng00031; NG00034 - https://www.niagads.org/datasets/ng00034). By calculating identity-by-descent using PLINK[53], we identified duplicates, which were excluded from the ADSP WES dataset for subsequent analyses.

1.1.1 UK Biobank study

The current study used data from the UK Biobank[54] (UKB; www.ukbiobank.ac.uk), a large population-based cohort that includes over 500,000 participants and aims to improve insight into a wide variety of health-related determinants and outcomes across the UK. Between 2006 and 2010, approximately 9.2 million invitations to participate in the study were sent to individuals aged 40-69 years who were registered with the National Health Service (NHS) and were living within 25 miles from one of the 22 study research centers. In total, 503,325 participants were recruited in the study, from which we used a subsample of individuals of European ancestry with available phenotypic and genotypic data (*M* age = 56.5, 54.0% female), described in more detail below. Besides phenotypic information obtained from the NHS registries and associated medical records, participants completed an in-person visit at one of the study research centers where extensive self-report data were collected by questionnaire in addition to anthropometric assessments, DNA collection from blood samples, and magnetic resonance imaging of body and brain. All participants provided written informed consent; the UKB received ethical approval from the National Research Ethics Service Committee North West-Haydock (reference 11/NW/0382), and all study procedures were in accordance with the World Medical Association for medical research. Access to the UK Biobank data was obtained under application number 16406.

1.2 UKB proxy phenotype

A proxy phenotype for Alzheimer's disease case-control status in UKB was assessed as part of the self-report questionnaire administered during the in-person assessment. Participants were asked to report whether their biological mother or father ever suffered from Alzheimer's disease/dementia, and to report each parent's current age (or age at death, if applicable). Of 376,113 individuals in our analytic subsample who completed these questions, a diagnosis was reported for 32,327 mothers (8.6%) and 17,014 fathers (4.5%), resulting in 47,793 participants (12.7%) with one or both parents affected. We created a proxy phenotype from these questions to index genetic risk for Alzheimer's based on parents' diagnoses. The phenotype was constructed as a linear count of the number of affected biological parents (0, 1, or 2). The contribution for each unaffected parent to this count was weighted by the parent's age/age at death to account for the fact that they may not yet have passed through the period of risk for this late-onset disease. Specifically, each affected parent contributed one full unit of "risk" to the count, while each unaffected parent contributed a proportion of one unit of "risk" inversely related to their age. This was calculated as the ratio of parent's age to age 100 (approximately the 95th percentile for life expectancy in developed countries), such that weight=(100-age)/100. The weight for unaffected parents was capped at 0.32, corresponding to a risk equivalent to that of the maximum population prevalence of AD.[55] The phenotype thus ranged approximately from

0 to 2, with values near zero when both parents were unaffected (lower for older parents and possible values below zero if both parents were over age 100) and values of two when both parents were affected. Participants who were uncertain or chose not to answer questions about either parent's disease status or age were excluded from the analyses, resulting in a final N=364,859.

Additional information on Alzheimer's disease risk was obtained from national medical records linked to participant data. This information pertained to the participants themselves (not their parents), and was extracted from hospital records obtained between 1996 and the present or from national death registries in the case of participants who passed away after initial enrolment in the study, as described in more detail in the UKB resources (http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=146641; http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=115559). Briefly, primary and secondary diagnoses from inpatient hospital stays and primary and secondary causes of death from death records were recorded using ICD-10 codes. Participants with a diagnosis of "Alzheimer's disease" (diseases of the nervous system chapter; code G30) or "Dementia in Alzheimer's disease" (mental and behavioral disorders chapter; code F00) from any record of a hospital stay or as a cause of death were treated as Alzheimer's cases as given the maximum possible "risk" score of 2, regardless of the affectation status of their parents. The reported rate of Alzheimer's in parents of cases (27.4%) was more than double that of non-cases (12.7%; $\chi^2(1)=71.7$, P=2.45E-17). There were 393 individuals in the analytic subsample classified as affected by these records; due to the small number of cases and the limited representativeness of these types of health records, we used this information to supplement the proxy parent phenotype rather than as a primary outcome. This information reduces the possibility of misclassification in the proxy phenotype method, and also allows us to evaluate the performance of the proxy phenotype.

1.3 Genome-wide association analysis

Except for IGAP (obtained summary statistics), we performed genome-wide association analyses for the ADSP, PGC-ALZ and UKB cohorts. For the UKB dataset, quality control and imputation procedures were slightly different, and therefore described separately in the sections below.

1.3.1a Quality control and imputation procedures for ADSP and PGC-ALZ datasets

Prior to individual quality control steps, all datasets were filtered on a max missingness of 5%. Individuals were excluded when identified as a low quality sample (individual call rate < 0.98), heterozygosity outlier (F +/-.20), gender mismatch (females: F >0.2, males: F < 0.2) when comparing phenotypic and genotypic data, population outlier (defined by principal component boundaries of 1000 Genomes European samples) or being related to another sample (PI_HAT > 0.2). Inclusion criteria for variants encompassed a call rate > 0.98, a case-control missingness difference < 0.02, a Hardy-Weinberg equilibrium *p*-value < $10x10^{-6}$ for controls (<$10x10^{-10}$ for cases) and a valid association *p*-value (excluding the variants with low allele frequencies).

Pre-imputation, the ADSP and PGC-ALZ datasets were checked for palindromic variants with allele frequency close to 0.5, incorrect reference allele definitions, false strand designation and extreme deviations from expected allele frequencies. Subsequently the ADSP and PGC-ALZ datasets were imputed with the 1000 Genomes Phase 3[56] reference panel. The reported SNPs all have a considerable imputation quality (depending on the dataset that was imputed, at least 72%

of lead SNPs have an INFO score > 0.9) and variants with a low allele frequency (MAF<0.01) were excluded, resulting in a total of 7508 individuals (4,343 cases and 3,165 controls) and 260,934 variants for the ADSP cohort and 17477 individuals (2,736 cases and 14,471 controls) and 9,629,492 variants for the PGC-ALZ cohort.

1.3.1b Quality control and imputation for UKB dataset

We used second-release genotype data that were made available by UKB in July 2017. Genotype data collection and processing are described by the UKB in a previous overview paper[57]. DNA was extracted from blood samples and genotyping was completed for 488,366 individuals on one of two Affymetrix genotyping arrays with custom content, the UK BiLEVE Axiom array (N=49,949) or UK Biobank Axiom array (N=438,417), covering 812,428 genetic markers common to both arrays. Of these, 488,377 individuals and 805,426 markers passed the genotype quality control checks conducted by UKB (see http://www.biorxiv.org/content/early/2017/07/20/166298 for details). Samples were excluded for low DNA concentration, call rate < 95%, excess heterozygosity, sex chromosome abnormality, or sample duplication. Variants were excluded if they exhibited poor clustering of allele calls, batch, plate, array, or sex effects, departures from HWE, or discordance between technical replicate samples.

After quality control, the samples were imputed to approximately 92 million SNPs using both the reference panel of the Haplotype Reference Consortium (HRC)[58] as well as a combined reference panel of the 1000 Genomes Project[56] and UK10K. As recommended by UKB, we removed variants that were not imputed on the HRC reference panel due to technical errors in the imputation process of the combined panel. We converted imputed variants to hard calls (certainty > 0.9), filtered by imputation quality (INFO score >0.9), and excluded multi-allelic SNPs, indels, SNPs without unique rsID, and SNPs with minor allele frequency (MAF) <0.0001, resulting in 10,847,151 SNPs available for analysis.

For the present study, we selected unrelated individuals of European ancestry. To empirically determine ancestry, we projected genetic principal components from known ancestral populations in the 1000 Genomes Project onto the UKB genotypes and assigned individuals to the continental ancestral superpopulation with the closest Mahalanobis distance.[59] Within-ancestry principal components were created using FlashPCA2[60] to correct for any residual population stratification within the European ancestry subset. Unrelated individuals (less than 3rd degree relatives, as indicated by genomic relatedness coefficients calculated by UKB) were selected by sequentially removing participants with the greatest number of relatives until no related pairs remained. After applying these filtering criteria and removing any participants with missing phenotypic or covariate data and participants who withdrew consent, 364,859 individuals remained for analysis in the UKB sample.

1.3.2 Single-marker association analysis

Genome-wide association analysis (GWAS) for the ADSP, PGC-ALZ and UKB datasets was performed in PLINK[53], using logistic regression for dichotomous phenotypes (cases versus controls for ADSP and PGC-ALZ cohorts), and linear regression for phenotypes analysed as continuous outcomes (by proxy parental AD phenotype for UKB cohort). For the ADSP and PGC-ALZ cohorts, association tests were adjusted for gender, batch (if applicable), and the first 4 principal components. Twenty principal components were calculated, and depending on the dataset being tested, additional principal components (on top of the standard inclusion of 4 PCAs) were added if significantly associated to the phenotype. Furthermore, for the PGC-ALZ cohorts age was included as a covariate. For 4,537 controls of the DemGene cohort, 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 dataset, age was not included as a covariate due to the enrichment for older controls (mean age cases = 73.1 years (SE=7.8); mean age controls = 86.1 years (SE=4.5)) in their collection procedures. Correcting for age in ADSP would remove a substantial part of genuine association signals (e.g. well-established *APOE* locus rs11556505 is strongly associated to AD (P=1.08x10^{-99}), which is lost when correcting for age (P=0.0054). For the UKB dataset, 12 components were included as covariates, as well as age, sex, genotyping array, and assessment centre. We used the genome-wide threshold for significance of P<5×10-8).

1.3.3 Multivariate genome-wide meta-analysis

Two meta-analyses were performed, including: 1) cohorts with case-control phenotypes (IGAP, ADSP and PGC-ALZ datasets), 2) all cohorts, also including the by proxy phenotype of UKB.

The per SNP test statistics is defined by

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

where w_i and Z_i are the squared root of the sample size and the test statistics of SNP k in cohort i, respectively. CTI is the cross-trait LD score intercept estimated by LDSC[14,61] using genome-wide summary statistics. This is equal to[61]

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

where N_i and N_j are the sample sizes of cohorts i and j and N_{sij} the number of samples overlapping between them, and ρ_{ij} 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 by using 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-value < 1e-323 to 1e-323 (the minimum >0 value in Python). The script for the multivariate GWAS is available from https://github.com/Kyoko-wtnb/mvGWAMA.

1.3.4 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 those other cohorts 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 (Methods section 1.14) 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 from https://github.com/Kyoko-wtnb/mvGWAMA.

1.4 Replication of meta-analysis result in an Icelandic sample

The study group included 6,593 Alzheimer's disease cases (4,923 of whom were chip-typed) and 174,289 controls (88,581 of whom were chip-typed). In 16% of patients, the diagnosis of Alzheimer's disease was established at the Memory Clinic of the University Hospital according to the criteria for definite, probable, or possible Alzheimer's disease of the National Institute of

Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA). In 77% of patients, the diagnosis has been registered according to the criteria for code 331.0 in ICD-9, or for F00 and G30 in ICD-10 in health records. Seven percent of the patients were identified in the Directorate of Health medication database as having been prescribed Donepezil (Aricept). The controls were drawn from various research projects at deCODE Genetics.

The study was approved by the National Bioethics Committee and the Icelandic Data Protection Authority. Written informed consent was obtained from all participants or their guardians before blood samples were drawn. All sample identifiers were encrypted in accordance with the regulations of the Icelandic Data Protection Authority.

Chip-typing and long-range phasing of 155,250 individuals was carried out as described previously.[20] Imputation of the variants found in 28,075 whole-genome sequenced individuals into the chip-typed individuals and 285,664 close relatives was performed as detailed earlier.[20] Association analysis was carried out using logistic regression with Alzheimer’s disease status as the response and genotype counts and a set of nuisance variables including sex, county of birth, and current age as predictors.[21] Correction for inflation of test statistics due to relatedness and population stratification was performed using the intercept estimate from LD score regression[14] (1.29).

1.5 Genomic risk loci definition

We used FUMA[27] 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 ($<5\times10^{-8}$) 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 <250kb apart). Borders of the genomic risk loci were defined by identifying all SNPs in LD ($r^2>0.6$) with one of the independent significant SNPs in the locus, and the region containing all these candidate SNPs was considered to be a single independent genomic risk locus. LD information was calculated using the UK Biobank 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[25]. This Bayesian model estimates a per-SNP 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**).

1.6 Conditional analysis

We performed conditional analysis with GCTA-COJO[62] 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 III summary statistics, using 10.000 randomly selected unrelated samples from the UKB dataset as a reference dataset to determine LD-patterns. Conditional analysis was run per chromosome with the default settings of the software.

1.7 Heritability and Genetic Correlation

LD score regression[14] was used to estimate clinical AD heritability and to calculate genetic correlations[61] between the case-control and proxy phenotypes using their post-quality control summary statistics. Pre-calculated LD scores from the 1000 Genomes European reference population were obtained from https://data.broadinstitute.org/alkesgroup/LDSCORE/. 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.

1.8 Polygenic risk scoring

We calculated polygenic scores (PGS) using two independent genotype datasets. First, 761 individuals (379 cases and 382 controls) from the ADDNeuroMed study[63] were included, using the same QC and imputation approach as for the other datasets with genotype-level data (see Method section 1.3.1a). Second, 1459 individuals (912 cases and 547 matched controls) from the TGEN study were assessed and their diagnostic status was confirmed via post-mortem neuropathology. Imputed SNPs in this sample were filtered based on INFO>0.9 and MAF>0.01. PGS were created using PLINK for the TGEN dataset and the PLINK-based software PRSice for the ADDNeuroMed dataset. In both samples, PGS 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 while preferentially selecting SNPs on lower GWAS *P*-values. Clumping was based on the effect size estimates of SNPs originating from the Phase III 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 III SNPs. PGS were calculated in both samples using the SNP effect size estimates from the Phase III 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 the same covariates as in each cohort-specific GWAS, compared to a linear model with GWAS covariates only. In the TGEN dataset, sensitivity, specificity, and area under the curve (AUC) of predicting confirmed case/control status were calculated, using the R package pROC and bootstrapping to obtain confidence intervals on the AUC estimate. Of note, approximately 3% of the TGEN sample overlapped with the IGAP cohort included in the meta-analysis; previous simulation work using PGS 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.[23]

1.9 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 (https://github.com/bulik/ldsc)[14]. Heritability enrichment was calculated as the proportion of heritability explained by a SNP category divided by the proportion of SNPs that are in that category. Partitioned heritability was computed by 28 functional annotation categories, by minor allele frequency (MAF) in six percentile bins and by 22 chromosomes. Annotations for binary categories of functional genomic characteristics (e.g. coding or regulatory regions) were obtained from the LD score website (https://github.com/bulik/ldsc). The Bonferroni-corrected significance threshold for 56 annotations was set at: $P<0.05/56=8.93\times10^{-4}$.

1.10 Functional Annotation of SNPs

Functional annotation of GWS SNPs implicated in the meta-analysis was performed using FUMA[27] v1.2.8 (http://fuma.ctglab.nl/). Functional consequences for these SNPs were obtained by matching SNPs' chromosome, base-pair position, and reference and alternative alleles to databases containing known functional annotations, including ANNOVAR[64] categories, Combined Annotation Dependent Depletion (CADD) scores[24], RegulomeDB[65] (RDB) scores, and chromatin states[66,67]. ANNOVAR annotates the functional consequence of SNPs on genes (e.g. intron, exon, intergenic). CADD scores predict how deleterious the effect of a SNP with higher scores referring to higher deleteriousness. A CADD score above 12.37 is the threshold to be potentially pathogenic[68]. The RegulomeDB score is a categorical score based on information from expression quantitative trait loci (eQTLs) and chromatin marks, ranging from 1a to 7 with lower scores indicating an increased likelihood of having a regulatory function. Scores are as follows: 1a=eQTL + Transciption Factor (TF) binding + matched TF motif + matched DNase Footprint + DNase peak; 1b=eQTL + TF binding + any motif + DNase Footprint + DNase peak; 1c=eQTL + TF binding + matched TF motif + DNase peak; 1d=eQTL + TF binding + any motif + DNase peak; 1e=eQTL + TF binding + matched TF motif; 1f=eQTL + TF binding / DNase peak; 2a=TF binding + matched TF motif + matched DNase Footprint + DNase peak; 2b=TF binding + any motif + DNase Footprint + DNase peak; 2c=TF binding + matched TF motif + DNase peak; 3a=TF binding + any motif + DNase peak; 3b=TF binding + matched TF motif; 4=TF binding + DNase peak; 5=TF binding or DNase peak; 6=other;7=None. The chromatin state represents the accessibility of genomic regions (every 200bp) with 15 categorical states predicted by a hidden Markov model based on 5 chromatin marks for 127 epigenomes in the Roadmap Epigenomics Project[39]. 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. The 15-core chromatin states as suggested by Roadmap are as follows: 1=Active Transcription Start Site (TSS); 2=Flanking Active TSS; 3=Transcription at gene 5' and 3'; 4=Strong transcription; 5= Weak Transcription; 6=Genic enhancers; 7=Enhancers; 8=Zinc finger genes & repeats; 9=Heterochromatic; 10=Bivalent/Poised TSS; 11=Flanking Bivalent/Poised TSS/Enh; 12=Bivalent Enhancer; 13=Repressed PolyComb; 14=Weak Repressed PolyComb; 15=Quiescent/Low.

1.11 Gene-mapping

Genome-wide significant loci obtained by GWAS were mapped to genes in FUMA[27] v1.2.8 using three strategies:

1. Positional mapping maps SNPs to genes based on physical distance (within a 10kb 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 (i.e. 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[69] v6, Blood eQTL browser[70], BIOS QTL browser[71]), and is based on cis-eQTLs which can map SNPs to

genes up to 1Mb apart. We used a false discovery rate (FDR) of 0.05 to define significant eQTL associations.

3. Chromatin interaction mapping was performed to map SNPs to genes when there is 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 Hi-C data of 14 tissue types from the study of Schmitt et al[72]. Since chromatin interactions are often defined in a certain resolution, such as 40kb, an interacting region can span multiple genes. If a SNPs 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[67] and the other region is located in a gene promoter region (250bp up and 500bp 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 FDR of 1×10^{-5} to define significant interactions, based on previous recommendations[44] modified to account for the differences in cell lines used here.

1.12 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 catalogued in BRAINEAC[28], CommonMind Consortium Portal[29] and xQTL Serve[30].

The BRAINEAC data consists of 134 neuropathologically confirmed control individuals of European descent from the UK Brain Expression Consortium[28], of which the eQTL data was obtained from http://www.braineac.org/. Overlapping eQTLs with a FDR < 0.05 are reported for the following brain regions: cerebellar cortex, frontal cortex, hippocampus, inferior olivary nucleus (sub-dissected from the medulla), occipital cortex, putamen (at the level of the anterior commissure), substantia nigra, temporal cortex, thalamus and intralobular white matter. The original source data does not provide the tested allele.

The CommonMind Consortium data consists of post-mortem brain samples of 467 Caucasian individuals (209 subjects with schizophrenia, 206 healthy controls, 52 subjects with bipolar or other affective/mood disorders). The eQTL data was obtained from https://www.synapse.org//#!Synapse:syn5585484, where we have used the version that corrects for SVA. The eQTLs are binned by FDR and therefore, nominal p-values are missing and the FDR p-values are reported as 0.009 or 0.049, to refer to < 0.01 and < 0.05, respectively.

The xQTL data is based on 494 post-mortem dorsolateral prefrontal cortex samples of a random set of older individuals, of which the eQTL data was obtained from http://mostafavilab.stat.ubc.ca/xqtl/ . Alignment of risk increasing allele and eQTL tested allele was not performed for this data source, since tested allele and signed statistics are not available in the original data source. For the mQTLs we assigned the nearest gene for each significantly associated methylation probe that overlapped one of the significant loci.

1.13 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 gene-based genome-wide association analysis (GWGAS) using MAGMA[35], followed by a meta-analysis. SNP-based P-values from the meta-analysis of the 3 genotype-array-based datasets were used as input for the first GWGAS, while the unimputed individual-level sequence data of ADSP was used as input for the second GWGAS. 18,233 protein-coding genes (each containing at least one SNP in the GWAS) from the NCBI 37.3 gene definitions were used as basis for GWGAS in MAGMA. Bonferroni correction was applied to correct for multiple testing ($P<2.74x10^{-6}$).

1.14 Gene-set analysis

Results from the GWGAS analyses were used to test for association in 7,086 predefined gene-sets of four types:

1. 6,994 curated gene-sets representing known biological and metabolic pathways derived from Gene Ontology (5917 gene-sets), Biocarta (217 gene-sets), KEGG (186 gene-sets), Reactome (674 gene-sets) catalogued by and obtained from the MsigDB version 6.1[73] (http://software.broadinstitute.org/gsea/msigdb/collections.jsp)
2. Gene expression values from 53 tissues obtained from GTEx[69], log2 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 [38]. Briefly, brain cell-type expression data

was 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 (UMI) counts for the given cell type by the summed mean UMI 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[74]. The value for each cell-type was calculated with the same method as explained in point 3 above.

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 was 0.05/7,087 gene-sets=7.06×10^{-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 all others and of *APOE* (a gene-set including all genes within genomic 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.

1.15 Cross-Trait Genetic Correlation

Genetic correlations (r_g) between AD and 41 phenotypes were computed using LD score regression[14], as described above, based on GWAS summary statistics obtained from publicly available databases (http://www.med.unc.edu/pgc/results-and-downloads; http:// ldsc.broadinstitute.org/; **Supplementary Table 26**). The Bonferroni-corrected significance threshold was 0.05/41 traits=1.22×10^{-3}.

1.16 Mendelian Randomisation

To infer credible causal associations between AD and traits that are genetically correlated with AD, we performed Generalised Summary-data based Mendelian Randomisation[37] (GSMR; http://cnsgenomics.com/software/gsmr/). 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 bi-directional 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=<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 (OR)[37] for a case-control trait. An OR of 2 can be interpreted as a doubled risk compared to the population prevalence of a binary trait for every SD increase in the exposure trait. For quantitative traits the b_{zx} and b_{zy} can be interpreted as a one standard deviation increase explained in the outcome trait for every SD 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.

Data availability

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

References

1. Prince M, Bryce R, Albanese E, Wimo A, Ribeiro W, Ferri CP. The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 2013; **9**(1): 63-75.e2.
2. Gatz M, Reynolds CA, Fratiglioni L, et al. Role of genes and environments for explaining Alzheimer disease. *Archives of general psychiatry* 2006; **63**(2): 168-74.
3. Cacace R, Sleegers K, Van Broeckhoven C. Molecular genetics of early-onset Alzheimer's disease revisited. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 2016; **12**(6): 733-48.
4. Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 2013; **45**(12): 1452-8.
5. Goate A, Chartier-Harlin MC, Mullan M, et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* 1991; **349**(6311): 704-6.
6. Sherrington R, Rogaev EI, Liang Y, et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* 1995; **375**(6534): 754-60.
7. Sherrington R, Froelich S, Sorbi S, et al. Alzheimer's disease associated with mutations in presenilin 2 is rare and variably penetrant. *Human molecular genetics* 1996; **5**(7): 985-8.
8. Karran E, Mercken M, De Strooper B. The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nature reviews Drug discovery* 2011; **10**(9): 698-712.
9. Jonsson T, Stefansson H, Steinberg S, et al. Variant of TREM2 associated with the risk of Alzheimer's disease. *The New England journal of medicine* 2013; **368**(2): 107-16.
10. Steinberg S, Stefansson H, Jonsson T, et al. Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. *Nature genetics* 2015; **47**(5): 445-7.
11. Liu CC, Liu CC, Kanekiyo T, Xu H, Bu G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature reviews Neurology* 2013; **9**(2): 106-18.
12. Liu JZ, Erlich Y, Pickrell JK. Case-control association mapping by proxy using family history of disease. *Nature genetics* 2017; **49**(3): 325-31.
13. Zheng J, Erzurumluoglu AM, Elsworth BL, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* 2017; **33**(2): 272-9.
14. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015; **47**(3): 291-5.
15. de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008; **17**(R2): R122-R8.
16. Desikan RS, Schork AJ, Wang Y, et al. Polygenic Overlap Between C-Reactive Protein, Plasma Lipids, and Alzheimer Disease. *Circulation* 2015; **131**(23): 2061-9.
17. Jun GR, Chung J, Mez J, et al. Transethnic genome-wide scan identifies novel Alzheimer's disease loci. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 2017; **13**(7): 727-38.

18. Guerreiro R, Wojtas A, Bras J, et al. TREM2 variants in Alzheimer's disease. *The New England journal of medicine* 2013; **368**(2): 117-27.

19. Sims R, van der Lee SJ, Naj AC, et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nature genetics* 2017; **49**(9): 1373-84.

20. Gudbjartsson DF, Helgason H, Gudjonsson SA, et al. Large-scale whole-genome sequencing of the Icelandic population. *Nature genetics* 2015; **47**(5): 435-44.

21. Steinthorsdottir V, Thorleifsson G, Sulem P, et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nature genetics* 2014; **46**(3): 294-8.

22. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics* 2015; **31**(9): 1466-8.

23. Valentina EP, J. MA, Matt H, John H. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. *Annals of Neurology* 2017; **82**(2): 311-4.

24. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014; **46**(3): 310-5.

25. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014; **511**(7510): 421-7.

26. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 2015; **47**(11): 1228-35.

27. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nature communications* 2017; **8**(1): 1826.

28. Ramasamy A, Trabzuni D, Guelfi S, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience* 2014; **17**(10): 1418-28.

29. Fromer M, Roussos P, Sieberts SK, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature neuroscience* 2016; **19**(11): 1442-53.

30. Ng B, White CC, Klein HU, et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nature neuroscience* 2017; **20**(10): 1418-26.

31. Gurses MS, Ural MN, Gulec MA, Akyol O, Akyol S. Pathophysiological Function of ADAMTS Enzymes on Molecular Mechanism of Alzheimer's Disease. *Aging and disease* 2016; **7**(4): 479-90.

32. Suh J, Choi SH, Romano DM, et al. ADAM10 missense mutations potentiate beta-amyloid accumulation by impairing prodomain chaperone function. *Neuron* 2013; **80**(2): 385-401.

33. Dries DR, Yu G. Assembly, maturation, and trafficking of the gamma-secretase complex in Alzheimer's disease. *Current Alzheimer research* 2008; **5**(2): 132-46.

34. Dumitriu A, Golji J, Labadorf AT, et al. Integrative analyses of proteomics and RNA transcriptomics implicate mitochondrial processes, protein folding pathways and GWAS loci in Parkinson disease. *BMC medical genomics* 2016; **9**: 5.

35. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 2015; **11**(4): e1004219.

36. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* 2014; **42**(Database issue): D1001-6.

37. Zhu Z, Zheng Z, Zhang F, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat Commun* 2018; **9**(1): 224.

38. Skene NG, Grant SG. Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment. *Frontiers in neuroscience* 2016; **10**: 16.

39. Kang J, Rivest S. Lipid metabolism and neuroinflammation in Alzheimer's disease: a role for liver X receptors. *Endocrine reviews* 2012; **33**(5): 715-46.

40. Loewendorf A, Fonteh A, Mg H, Me C. Inflammation in Alzheimer’s Disease: Cross-talk between Lipids and Innate Immune Cells of the Brain; 2015.

41. Stern Y. Cognitive reserve in ageing and Alzheimer's disease. *The Lancet Neurology* 2012; **11**(11): 1006-12.

42. Satizabal C, Beiser AS, Seshadri S. Incidence of Dementia over Three Decades in the Framingham Heart Study. *The New England journal of medicine* 2016; **375**(1): 93-4.

43. Adams HH, Hibar DP, Chouraki V, et al. Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nature neuroscience* 2016; **19**(12): 1569-82.

44. Ikram MA, Fornage M, Smith AV, et al. Common variants at 6q22 and 17q21 are associated with intracranial volume. *Nature genetics* 2012; **44**(5): 539-44.

45. Graves AB, Mortimer JA, Larson EB, Wenzlow A, Bowen JD, McCormick WC. Head circumference as a measure of cognitive reserve. Association with severity of impairment in Alzheimer's disease. *The British journal of psychiatry : the journal of mental science* 1996; **169**(1): 86-92.

46. Abbott RD, White LR, Ross GW, et al. Height as a marker of childhood development and late-life cognitive function: the Honolulu-Asia Aging Study. *Pediatrics* 1998; **102**(3 Pt 1): 602-9.

47. Giuffrida ML, Tomasello F, Caraci F, Chiechio S, Nicoletti F, Copani A. Beta-amyloid monomer and insulin/IGF-1 signaling in Alzheimer's disease. *Molecular neurobiology* 2012; **46**(3): 605-13.

48. Magnusson PK, Almqvist C, Rahman I, et al. The Swedish Twin Registry: establishment of a biobank and other recent developments. *Twin research and human genetics : the official journal of the International Society for Twin Studies* 2013; **16**(1): 317-29.

49. Finkel D, Pedersen NL. Processing Speed and Longitudinal Trajectories of Change for Cognitive Abilities: The Swedish Adoption/Twin Study of Aging. *Aging, Neuropsychology, and Cognition* 2004; **11**(2-3): 325-45.

50. Gold CH, Malmberg B, McClearn GE, Pedersen NL, Berg S. Gender and health: a study of older unlike-sex twins. *J Gerontol B Psychol Sci Soc Sci* 2002; **57**(3): S168-76.

51. Gatz M, Fratiglioni L, Johansson B, et al. Complete ascertainment of dementia in the Swedish Twin Registry: the HARMONY study. *Neurobiology of aging* 2005; **26**(4): 439-47.

52. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 1984; **34**(7): 939-44.

53. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015; **4**: 7.

54. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; **12**(3): e1001779.

55. Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology* 2013; **80**(19): 1778-83.

56. The Genomes Project C. A global reference for human genetic variation. *Nature* 2015; **526**: 68.

57. Davies G, Marioni RE, Liewald DC, et al. Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112 151). *Mol Psychiatry* 2016; **21**(6): 758-67.

58. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; **48**(10): 1279-83.

59. Peterson RE, Edwards AC, Bacanu SA, Dick DM, Kendler KS, Webb BT. The utility of empirically assigning ancestry groups in cross-population genetic studies of addiction. *Am J Addict* 2017; **26**(5): 494-501.

60. Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics (Oxford, England)* 2017; **33**(17): 2776-8.

61. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015; **47**(11): 1236-41.

62. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics* 2012; **44**(4): 369-75, s1-3.

63. Lovestone S, Francis P, Kloszewska I, et al. AddNeuroMed--the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Annals of the New York Academy of Sciences* 2009; **1180**: 36-46.

64. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010; **38**(16): e164.

65. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 2012; **22**(9): 1790-7.

66. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012; **9**(3): 215-6.

67. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015; **518**(7539): 317-30.

68. Amendola LM, Dorschner MO, Robertson PD, et al. Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome research* 2015; **25**(3): 305-15.

69. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, NY)* 2015; **348**(6235): 648-60.

70. Westra HJ, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* 2013; **45**(10): 1238-43.

71. Zhernakova DV, Deelen P, Vermaat M, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet* 2017; **49**(1): 139-45.

72. Schmitt AD, Hu M, Jung I, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell reports* 2016; **17**(8): 2042-59.

73. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 2005; **102**(43): 15545-50.

74. Habib N, Avraham-Davidi I, Basu A, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature methods* 2017; **14**(10): 955-8.