Genome wide association neural networks (GWANN) identify novel genes linked to family history of Alzheimer's disease in the UK Biobank

Upamanyu Ghose * [1, 2]	upamanyu.ghose@psych.ox.ac.uk
William Sproviero [1]	william.sproviero@psych.ox.ac.uk
Laura Winchester [1]	laura.winchester@psych.ox.ac.uk
Marco Fernandes [1]	marco.fernandes@psych.ox.ac.uk
Danielle Newby [3, 1]	danielle.newby@ndorms.ox.ac.uk
Brittany S. Ulm [2, 7]	brittany.ulm@ndph.ox.ac.uk
Liu Shi [1]	liu.shi@psych.ox.ac.uk
Qiang Liu [1]	qiang.liu@psych.ox.ac.uk
Cassandra Adams [2, 4]	cassandra.adams@cmd.ox.ac.uk
Ashwag Albukhari [2, 5]	aalbukhari@kau.edu.sa
Majid Almansouri [2, 6]	malmansouri@kau.edu.sa
Hani Choudhry [2, 7]	hchoudhry@kau.edu.sa
Cornelia van Duijn [2, 7, 8]	cornelia.vanduijn@ndph.ox.ac.uk
Alejo Nevado-Holgado [1, 2, 7]	alejo.nevado-holgado@psych.ox.ac.uk

[1] Department of Psychiatry, University of Oxford, Oxford OX3 7JX, United Kingdom.
[2] Centre for Artificial Intelligence in Precision Medicines.
[3] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford OX3 7HE, United Kingdom.
[4] Centre for Medicines Discovery, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7BN, United Kingdom.
[5] Biochemistry Department, Faculty of Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia.
[6] Clinical Biochemistry Department, Faculty of Medicine, King Abdulaziz University, Jeddah 21589, Saudi Arabia.
[7] Big Data Institute, University of Oxford, Oxford OX3 7LF, United Kingdom.
[8] Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, United Kingdom.
* Corresponding Author

Abstract

Augmenting traditional genome wide association studies (GWAS) with advanced machine learning algorithms can allow the detection of novel signals in available cohorts by providing complementary approaches to the existing methods. We introduce “Genome wide association neural networks (GWANN)”, a novel approach that uses neural networks (NNs) to account for nonlinear and SNP-SNP interaction effects. We applied GWANN to family history of Alzheimer’s disease (AD) in the UK Biobank. Our method identified 26 known AD genes, 2 target nominations and 67 novel genes, and validated the results against brain eQTLs, AD phenotype associations, biological pathways, disease associations and differentially expressed gene sets in the AD brain. Some drugs targeting novel GWANN hits are currently in clinical trials for AD. Applying NNs for GWAS, alongside existing methods, illustrates their potential to complement existing algorithms and methods, and enable the discovery of novel and tractable targets for AD.

KEYWORDS: Alzheimer’s disease, neural networks, artificial intelligence, GWAS

ABBREVIATIONS: NN - Neural Network; GWANN - Genome Wide Association Neural Networks; UKB - UK Biobank; PPI - Protein-Protein Interaction; GSEA - Gene Set Enrichment Analysis; ORA - Over Representation Analysis; CDF - Cumulative Distribution Function; GO - Gene Ontology

1 Introduction

Alzheimer's disease (AD) affects approximately 30 million people in the world, making it the most common form of dementia[1]. It is characterised by the build-up of Aβ and tau proteins in the brain, leading to neuronal death and impaired cognitive function[2]. In the last 10 years, genome wide association studies (GWAS) have revolutionised our understanding of the inherited basis of disease and they have been critical in identifying multiple risk loci and novel disease pathways associated with AD involving the microglia and lysosome[3]. However, the classical GWAS analysis depends on sample size, and despite the number of SNPs identified until today, they still only explain a small fraction of the heritability of the disease[4]. Although gene-based methods were developed to identify the joint effects of rare variants[5], or gene-level analysis from GWAS summary data[6], there are currently no methods to perform gene-based studies for common variants.

Along with the modern availability of large datasets[7–9], to complement and enhance current GWAS methods, we propose to use an approach based on machine learning to shed light on more complex patterns in genomic mechanisms involving gene interactions and non-linear relationships. Machine learning methods, more particularly Neural Networks (NNs), have been instrumental in the advancement of multiple engineering industries due to their efficacy in analysing complex data patterns[10], especially where large amounts of data are available. Compared to the success of classical GWAS, the successes of NN in gene discovery has been limited. NNs have recently been employed and tested on various complex traits and diseases including eye colour and schizophrenia[11]. Our aim was to develop NNs specialised to perform a gene-level GWAS using SNP data available in the UK Biobank (UKB)[12]. Our method is gene-based and considers groups of SNPs within and around each gene in the genome to establish the association of the gene with the phenotype of interest. In this paper we demonstrate the application of our new GWANN method as a complementary method to existing GWAS methods, to identify associations with familial

history of Alzheimer's disease/dementia, a proxy that has been successfully used to identify new genes for AD in the UKB[7]. We present the genetic associations to family history of AD found by the method, and systematically validate the results against brain eQTLs, AD and related phenotype associations, biological pathways, disease associations and differentially expressed gene sets in the AD brain.

2 Results

2.1 Identification of novel genes related to AD using GWANN

We applied GWANN to 26,133 cases and controls of maternal history of AD, and 12,680 cases and controls of paternal history of AD. The analysis was run at a gene-level, where the SNPs within a gene and in the flanking 2500 bp region were mapped to the gene. Each gene was divided into windows containing a maximum of 50 SNPs, and a separate model (Figure 1) was trained for each window. In addition to the SNPs, age (field 21003), sex (field 31), the first six genetic principal components (PCs) obtained from UKB variables (field 22009), education qualification (field 6138), and a binary variable if this data was not available (1 if missing and 0 if not) were used as covariates.

GWANN identified 69 genes significantly associated with maternal history, 29 genes associated with paternal history, and 95 genes in the parental (i.e., maternal and paternal) meta-analysis (Figure 3, Supplementary Table 1), after multiple-comparison correction based on a Bonferroni threshold of P value<6.82×10^{-7} (Figure 2, Supplementary Table 2). Table 1 lists the genes with their P values. The flanking region of 2500 bp around each gene caused an overlap between windows of *GEMIN7* and *PPP1R37*, and *APOE* and *TOMM40*. However, there was no overlap between the significant windows in the first case, and in the second case, *TOMM40* had a significant non-overlapping window in addition to the significant window that overlapped with *APOE*. Additionally, for the significantly associated genes, we used the gradient of the inputs with respect to the loss function of the NNs to identify the top input features (SNPs or covariate encodings) which had the largest influence on the prediction of the NN. The top three features were among the SNPs within the windows of the hit genes in both, maternal and paternal history analyses, except *ARHGAP24, PITPNM2, UBE3A, CRISPLD1 and GEMIN7*, in the paternal history analysis, where the NNs gave more importance to some of the covariate encodings (Supplementary

Table 2). Apart from the *APOE* locus, significant gene windows in *LRP1B* and *PARD3B*; *CNTN4* and *ROBO1*; *STAC* and *CADM2*; *TYW1B, EPHA1-AS1* and *MAGI2*, were in linkage disequilibrium (LD, $r^2 \geq 0.8$) with each other (Supplementary Table 3). The significant genes were further than 500 kbp from each other, except for the *APOE* locus on chromosome 19, and *PTPRN2* and *VIPR2* on chromosome 7 (440 kbp). Supplementary File 2 contains the locus zoom plots for the GWANN hits, which show a neighbourhood of 500 kbp around the hit genes.

A standard SNP-level GWAS performed using PLINK 2.0 with a logistic regression model, on the same set of SNPs, covariates, and individuals as used in GWANN, identified 13 genes for maternal history, four genes for paternal history, and 14 genes in the meta-analysis, all of which were in the *APOE* locus, and previously identified in the UKB[7]. In this analysis, the SNPs were mapped to genes using the same mapping as used for GWANN. Five of these genes (*APOE* locus – *APOE, APOC1, TOMM40, BCAM* and *EXOC3L2*) overlapped with the findings of GWANN.

When compared with other GWAS on familial history of AD, AD diagnosis, and AD-related traits listed in the GWAS catalogue[3], there was an overlap of 26 genes with the GWANN hits. Twelve of these have been previously associated with AD (*ANK3, APH1B, GLIS3, EPHA1-AS1/EPHA1, WWOX* in addition to seven genes in the *APOE* region). Seven others are listed in the GWAS catalogue as suggestive associations (*P value*$<1x10^{-5}$) with AD (*EXOC4, NRXN1, NRXN3, PAX5, CADM2, LUZP2* and *RBMS3*). *DCC*, *CADM2*, *ARHGAP15 and PITPNM2* were identified in a recent analysis on pleiotropic predisposition to AD and educational attainment[13], and *EFNA5* and *LRP1B* as suggestive associations. *DAB1* has previously been associated with psychosis in AD[14], and *MACROD2* with neurofibrillary tangles[15].

To further understand the 69 GWANN hits that have not previously been identified in a GWAS on AD or AD-related traits, we obtained information about them from the Agora AD knowledge portal, which contains information about AD associated genes, collected by the National Institute on Aging's Accelerating Medicines Partnership in Alzheimer's Disease (AMP-AD) consortium as well as other research teams (https://agora.adknowledgeportal.org). *BDNF* and *ROBO1* have been nominated as potential targets for AD. Among the other novel genes identified by our method, *RNF182* and *SAMD4A* have been reported to be significantly associated with Braak stage (neurofibrillary tangles) and CERAD (measure of neuritic plaques); and *CPNE5, SAMD4A, ATP1B3, MAGI2* and *ARSG* associated with COGDX (clinical consensus diagnosis of cognitive status at time of death) (Table 2). Furthermore, from this set of 69 novel genes, 39 genes (Supplementary Figure 1) have associations in the GWAS catalogue with at least one disorder under the category of “nervous system disease” (EFO:0000618), comprising non-neoplastic or neoplastic disorders that affect the brain, spinal cord, or peripheral nerves.

2.2 Enriched protein-protein interaction (PPI) network of GWANN hits

The PPI network of the 95 associated genes identified post meta-analysis was significantly enriched (Figure 4a). For the 95 possible proteins, the expected number of connections in their PPI network by chance would be 23. However, the network of the GWANN hits show 72 connections, thereby being significantly enriched (*P value=3.33x10*$^{-16}$). On retaining only the 69 genes that have not been identified in previous GWAS on AD or AD-related traits, the PPI network was significantly enriched when the STRING interaction score was set at “low confidence (0.15)” (*P value= 3.13x10*$^{-12}$), but not enriched for higher thresholds of interaction score (*P value=0.172*). There were 93 indirect connections (simple paths) between the novel genes in the PPI network shown in Figure 4a, and 66 of these connections had a known AD gene as an intermediate node. The green dotted lines in Figure 4b shows the indirect

connections. The loss of interactions after removing the known genes from the network resulted in the drop in significance.

2.3 Enrichment of GWANN hits using transcriptomic data from AD post mortem brains

Using the results of a recent large meta-analysis of AD brain transcriptomic data by Patel et al.[16], we verified whether the results of GWANN show significant enrichment in the set of differentially expressed genes (DEGs) in the cerebellum, frontal lobe, parietal lobe, and temporal lobe. The authors reported results for DEGs in different groups of patients, denoting these as S_1=AD (patients with AD) and S_2=non-AD (i.e., patients with mental disorders different from AD). We ran our analysis for four different sets of DEGs, (i) S_1, (ii) S_2, (iii) $S_1 - S_2$ (patients with only AD and no other mental disorder), and (iv) $S_1 \cap S_2$ (patients with any mental disorder). For each set, the DEGs in the cerebellum (*P value=*$1.37x10^{-3}$, $1.04x10^{-3}$, $5.49x10^{-3}$, $4.13x10^{-3}$) and parietal lobe ($8.77x10^{-7}$, $7.31x10^{-11}$, $5.82x10^{-4}$, $4.67x10^{-6}$), and for sets i, ii and iii, the frontal lobe ($6.38x10^{-3}$, $6.46x10^{-5}$, $1.10x10^{-2}$) were enriched at a significance level of $1.25x10^{-2}$ (=*.05/4*). For the temporal lobe, there were no DEGs reported by the authors for set iv, and there was no enrichment in sets i, ii or iii (Table 3).

Another study by Patel et al. compared gene expression between controls, asymptomatic AD cases, and AD cases (GSE118553)[17]. They reported DEGs in the cerebellum, entorhinal cortex, frontal cortex, and temporal cortex, for three different conditions — (i) AD vs controls, (ii) asymptomatic AD vs AD, and (iii) asymptomatic AD vs controls. DEGs in the entorhinal cortex for AD vs controls, and asymptomatic AD vs AD (*P value=* $3.11x10^{-5}$, $9.39x10^{-5}$), and DEGs for AD vs controls in the temporal cortex ($1.49x10^{-3}$), showed enrichment with the GWANN summary statistics, at a significance level of $1.25x10^{-2}$ (=*.05/4*). In the case of asymptomatic AD vs controls, none of the brain regions showed enrichment. Supplementary

Figures 2 and 3, respectively, contain the cumulative distribution plots used in the enrichment of each dataset.

Furthermore, a total of 69 of the 95 associated genes (72.6%) were significantly enriched with eQTLs in brain regions, and 51 genes showed differential RNA expression in post-mortem AD brains (https://agora.adknowledgeportal.org) (Supplementary Table 4).

2.4 Enriched biological pathways, GO terms, diseases, and tissues, using GWANN summary statistics and hit genes

The enriched pathways include statin inhibition of cholesterol production (*FDR=3.69x10*$^{-4}$); regulation of commissural axon pathfinding by *SLIT* and *ROBO* (*3.69x10*$^{-3}$); *NTRK2* signalling through *RAS* and *CDK5* (*5.41x10*$^{-3}$*, 6.03x10*$^{-3}$); *RUNX1* regulation of transcription of genes involved in BCR signalling (*6.32x10*$^{-3}$); disruption of postsynaptic signalling by copy number variations (CNV) (*8.38x10*$^{-3}$); and Reelin signalling (*8.99x10*$^{-3}$). A complete list of pathways enriched in the gene set enrichment analysis (GSEA) for maternal history, paternal history and parental meta-analysis can be found in Supplementary Table 5.

In the over-representation analysis (ORA) (Figure 5a) using gene ontology (GO) terms, 56% (Figure 5b) of the terms were related with the cellular response to nerve growth factor stimulus (GO:1990090, corrected *P value=2.54x10*$^{-4}$). Other significant terms included Rho protein signal transduction (GO:0035023, *3.35x10*$^{-3}$); neuron recognition (GO:0008038, *2.19x10*$^{-2}$); amyloid-beta metabolic process (GO:0050435, *1.52x10*$^{-2}$); cellular component of the Schaffer collateral - CA1 synapse (GO:0098685, *1.47x10*$^{-2}$); Golgi to plasma membrane transport (GO:0006893, *5.35x10*$^{-3}$); and ephrin receptor signalling pathway (GO:0048013, *2.73x10*$^{-3}$). The results for GO enrichment of maternal and paternal histories can be found in Supplementary Figure 4 and Supplementary Table 7.

To further ascertain possible disease associations, we also calculated enrichment in DisGeNET (Table 4, Figure 5c), obtaining significance in multiple disorders including AD ($2.42x10^{-5}$), schizophrenia ($1.38x10^{-6}$), autistic disorder ($6.67x10^{-7}$), bipolar disorder ($3.54x10^{-4}$) and major depressive disorder ($4.49x10^{-3}$). *APOE, ANK3, BDNF* and *NTRK2* were common to all five disorders. Among the novel GWANN genes (those not detected in a previous GWAS on AD or AD-related traits), there was an overlap of (i) seven genes between AD and schizophrenia (*UBE3A, BDNF, NTRK2, SAMD4A, SERPINI1, CTNNA3, FLNB*); (ii) six between AD and bipolar disorder (*BDNF, NFIA, ASIC2, NTRK2, BCR, SORCS2*); (iii) five between AD and autistic disorder (*UBE3A, BDNF, ASIC2, NTRK2, CTNNA3*); and (iv) six between AD and depressive disorder (*BDNF, ASIC2, ARHGAP24, NTRK2, BCR, CTNNA3*). Supplementary Table 6 contains the list of GWANN genes overlapping with all the enriched DisGeNET disorders. The GSEA also resulted in enrichment of the pathways for Alzheimer's disease ($2.98x10^{-3}$) and amyotrophic lateral sclerosis ($3.76x10^{-4}$), amongst all KEGG pathways.

The most enriched tissue groups were the brain (*P value*$=1.16x10^{-15}$), blood vessel ($1x10^{-13}$) and nerve ($5.98x10^{-9}$). Figure 5d shows the enriched tissues in the GTEx v8 54 tissue types.

2.5 Potential of novel targets for AD drug discovery

Seventeen novel hits were reported as tractable targets for drug discovery (Supplementary Table 8). *NTRK2, FSHR, BCR, PDE1C* and *EPHA6*, are five novel GWANN hits that have known approved drugs (Table 4). Among these, *BCR* has 2 drugs — nilotinib and dasatinib — associated with it, one of which has undergone, and the other is currently undergoing clinical trials for AD. From the list of enriched pathways, *NTRK2* signalling plays an important role in synaptic transmission and neuronal development (R-HSA-9032845 and R-HSA-9026519) and may be a potential therapeutic target for AD. *DCC*, another GWANN hit that was also recently associated with pleiotropic predisposition to Alzheimer's disease and

educational attainment[13], is not currently associated with any approved drug but received the highest tractability probability of 98.47%. Given its importance in axon guidance (enriched pathway R-HSA-428542), it could potentially be another therapeutic target to explore.

3 Discussion

We developed GWANN and applied it to identify genes associated with maternal and paternal histories of AD using data from the UKB, followed by a parental meta-analysis. In doing so, we were able to identify 26 known AD genes, when compared with previous GWAS on AD or AD-related traits, two AD target nominations, and 67 novel genes. The post-hoc enrichment analyses showed enriched biological and disease pathways relevant to AD and neurodegeneration, and other neuropsychiatric diseases. Several GWANN hits were also identified as drug targets.

3.1 GWANN identifies hits associated with AD diagnosis

Of the genes that overlap between GWANN's discovery and discovery from earlier GWAS, most of them have also been identified previously using familial history data in the UKB. However, GWANN also identified previously associated or suggestively associated genes from GWAS performed on cohorts of diagnosed AD patients that have not been identified in the UKB. *GLIS3* was previously identified in a GWAS performed on cerebrospinal fluid tau levels[18]; *WWOX* was previously identified in the IGAP data in an analysis that did not include UKB[19]. *IMMP2L*, although not identified in a previous GWAS, was previously identified in an analysis on CNVs associated with AD[20]. Among the suggestive associations, other than the genes possibly linked to AD diagnosis, *EFNA5* was suggestively associated with psychosis in AD[14], and *MACROD2* with neurofibrillary tangles[15]. *ANK3*, a novel discovery in the GWAS conducted by Bellenguez et al. on the European Alzheimer's Disease Biobank (EADB)[21], was a GWANN hit. The ability of GWANN to identify these genes from the UKB data suggests the presence of subtle patterns in the data that were possibly missed by the traditional linear models but identified by our method.

3.2 Role of hit genes and enriched biological pathways in neurodegeneration and AD

The enriched pathways of the novel hits and known suggestive hits play important roles in the pathogenesis of AD. For example, axon guidance molecules such as netrin-1 play an important role in the regulation of Aβ levels and Reelin levels[22]. Two enriched pathways, R-HSA-428542 and R-HSA-8866376 are branched under the axon guidance pathway. *DCC,* a GWANN hit, is one of the main receptors of netrin-1 and an integral component of the enriched pathway R-HSA-428542. Through (i) its interaction with *ROBO1* (GWANN hit and AD target nomination) in the presence of *SLIT*, plays an important role in controlling commissural axon growth; and (ii) its interaction with *APP,* mediates axon guidance by enhancing intracellular signalling, and may have a role in the negative regulation of Aβ formation[23]. Furthermore, the phosphorylation of *DAB1*, another GWANN hit, plays a role in Reelin signalling (R-HSA-8866376). The loss of Reelin function in humans has previously been associated with AD[24].

Another GWANN hit with previous suggestive association to AD, *PAX5*, is a transcription co-factor along with the *RUNX1* complex that plays an important role in BCR signalling and B cell development (enriched pathway R-HSA-8939245). The role of B cells in the pathogenesis of central nervous system diseases is well established, and B cell depleting therapies have shown success in patients with disorders such as multiple sclerosis[25]. A recent study in mice has also suggested the involvement of B cells in neurodegenerative diseases like AD by causing immunoglobulin deposits around Aβ plaques[26].

AD is characterised by synaptic failure[2], and disruption of postsynaptic signalling by CNVs (WP4875) was among the enriched pathways in the GSEA. Two GWANN hits, *NRXN1* and *NRXN3,* belong to the neurexin family of proteins and contribute significantly to

synaptogenesis[27]. The disruption of the functioning of these genes due to CNVs affect synapse formation and neurodevelopment[27]. They have also previously been associated with autism[28] and schizophrenia[29]. Recent studies suggest the intersection between the aetiologies of AD and schizophrenia[30,31]. Another integral component of synaptic transmission is controlled by neurotrophins like *BDNF* (GWANN hit and nominated AD target), which play an important part in the survival, differentiation, and plasticity of neurons. Low levels of *BDNF* mRNA in patients with AD and mild cognitive impairment is an experimentally reproduced finding, and BDNF treatments in rodent and primate models of AD have shown success previously[2]. *NTRK2* (or *TRKB*), a novel GWANN hit, is the main receptor that *BDNF* binds to; *NTRK2* signalling plays a key role in neuronal development and possibly affects hippocampal long-term potentiation (R-HSA-9620244)[2] through *CDK5* catalytic activity (R-HSA-9032845), two pathways that were also enriched in the GSEA.

3.3 Repurposing known drugs of novel hits for AD

A large proportion of the known drugs associated with the novel GWANN hits belong to a broader group of drugs known as tyrosine kinase inhibitors (TKI). Drugs like nilotinib and dasatinib, among other TKIs, are being tested to be repurposed for AD due to their ability to reduce tau hyperphosphorylation and reverse Aβ-induced synaptic dysfunction and synapse loss in mouse models of AD-related pathology[32]. A previous study using mouse models also identified the effect of TKIs in reducing Aβ levels and astrocyte and dendritic cell number after treatment with nilotinib and bosutinib, two drugs associated with the novel GWANN hit *NTRK2*. Other than TKIs, dipyridamole, a drug associated with the novel hit *PDE1C*, was previously shown to prevent Aβ-induced microglial inflammation, thereby making it a possible therapeutic intervention for AD[33]. Pentoxifylline, a methylxanthine associated with *PDE1C*, has been suggested to improve cognitive function in patients with vascular dementia[34], reduce the odds of occurrence of Parkinson's disease[35], and reduce Aβ levels[35]. This further supports the quality of the novel GWANN hits and (i) suggests the possibility of

repurposing the existing drugs for AD, and (ii) provides evidence that several of these drugs are already being explored for the treatment of AD.

3.4 Limitations and considerations

Some AD hits such as *CLU*, *BIN1* and *CR1* were not identified by GWANN as genes that reached the genome wide significance level. However, when we observe the neighbourhood of these genes (Supplementary File 3, Supplementary Figure 5), we see that *BIN1* crossed the nominal significance threshold *(P value =* $3.2x10^{-6}$); the *CR1/CR2* locus was close to nominal significance ($4.39x10^{-5}$); and *PBK*, a gene in LD ($r^2 \geq 0.25$) with *CLU*, also had a P value close to nominal significance ($6.25x10^{-5}$). Additionally, *GPR137C*, a gene in LD ($r^2 \geq 0.8$) with *FERMT2*, a hit identified previously in the UKB[7], and in the analysis by Bellenguez et al.[21], was a GWANN hit (Supplementary File 3). To further explore this, we compared the distribution of minor allele frequencies (MAFs) of significant SNPs in previous GWAS, and SNPs in gene windows identified as significant by GWANN. The significant SNPs in the genes not identified by GWANN had a different distribution (higher MAFs) as compared to the known and novel genes identified by GWANN (Supplementary Figure 6, Supplementary File 4), thereby suggesting GWANN to be a complementary method alongside existing GWAS methods. Furthermore, some of the lead SNPs associated with the genes missed by GWAS, were not analysed by GWANN because of limiting the flanking region around genes to 2500 bp, due to computational feasibility. Hence, the difference in sample sizes, cohorts, and SNPs used by GWANN, along with the difference in analysis model, as compared to previous GWAS could be the factors contributing to the inability of these loci to reach genome wide significance in this analysis. In future analyses, we plan to test the method using larger sample sizes (larger ratio of controls:cases, instead of a 1:1 ratio), and to modify the method to accommodate larger genomic windows. Given that the method involves training a large number of NNs, we acknowledge the fact that it requires a fairly large amount of computational time and resources. However, we believe that with the rapidly

advancing field of ML and AI, it is possible to further optimise this method to achieve a speed up in computational time and reduction in resource usage. For example, the method of “knowledge distillation” uses the concept of a larger teacher NN teaching a much smaller student NN to learn the same task, which allows a significant speed up in time[36]. Secondly, NNs are more difficult to interpret as compared to traditional linear models, thereby rendering them as “black boxes". In this work we explored the gradients of the NNs to identify the SNPs that were most important in driving the prediction, but more methods to increase their interpretability and identify interactive effects would be beneficial in further understanding the hits identified, and those missed by GWANN. Finally, we acknowledge the fact that the discovery in this work warrants replication in different cohorts having diagnosis of AD. Our method can potentially be extended to other data sources, as well as be applied to other diseases or groups of diseases to understand the common associations, if any, between them.

3.5 Conclusion

We applied our method to family history of AD using data from the UKB and introduced a new method to complement the success of existing GWAS methods. GWANN identified novel genes associated with family history of AD, which led to the enrichment of new pathways such as *NTRK2* signalling through *CDK5*, and identification of possible drugs associated with novel genes that are currently being or could be repurposed for AD. This holds the potential to augment the effort of existing methods in understanding the pathogenesis of AD and other diseases.

4 Methods and Materials

4.1 Population

We utilised data from the UK Biobank (UKB) (http://www.ukbiobank.ac.uk). The data comprises health, cognitive and genetic data collected from ~500,000 individuals aged between 37 and 73 years from the United Kingdom at the study baseline (2006–2010)[12,37]. We used imputed SNP genotype data as input to GWANN. UKB genotyping was conducted by Affymetrix using a BiLEVE Axiom array for 49,950 participants and on a further updated using an Affymetrix Axiom array for the remaining 438,427 individuals in this study, based on the first array (95% marker content shared). The released genotyped data contained 805,426 markers on 488,377 individuals. Information on the genotyping process is available on the UKB website (http://www.ukbiobank.ac.uk/scientists-3/genetic-data)[37]. Genotype imputation was performed combining the UK 10K haplotype and Haplotype Reference Consortium (HRC) as reference panels[38]. A number of individuals (*n=856*) either with inconsistencies between their genetic predicted and reported sex, or abnormal number of sex-chromosomes were removed. In addition, 968 outliers were identified based on heterozygosity and missingness, and removed. The dataset was further limited to only individuals of “White British” descent resulting in 409,703 remaining individuals. A genetic relationship matrix along with genome-wide complex trait analysis was used to identify 131,818 individuals with relatives within the dataset, using a relationship threshold of 0.025. Only one person from each pair of related individuals was retained. Only biallelic SNPs with MAF > 1% and imputation quality info score > 0.8 were retained for the analysis, and all indels and multi-allelic SNPs were dropped.

4.2 Definition of cases and controls

The case groups consisted of individuals with maternal and paternal histories of AD, some of whom also had AD diagnosis. Individuals with diagnosed AD but no familial history of AD, and those with other neurological disorders[39] were removed from the control groups. Supplementary Table 9 contains a list of all the neurological disorders along with the UKB fields used to determine the presence of the disorders. Those with both maternal and paternal histories were dropped to avoid any overlap between the two datasets. We divided the entire range of ages into three groups (age-group1: 38-52, age-group2: 53-61, age-group3: 62-73 years), and paired them with the two possibilities of the individuals' sex (male and female) to obtain six broad groups — (age-group1, male), (age-group1, female) etc. To deal with the problem of imbalanced classes which affects the training of machine learning algorithms, we opted to randomly downsample the controls, while balancing for the six groups, to match the number of cases and obtain a 1:1 ratio. The sets of individuals in the datasets for paternal history and maternal history analyses were not overlapping. After these steps, maternal history had 26,133 cases and controls, and paternal history had 12,680 cases and controls.

4.3 Neural network model

GWANN follows an architecture with 2 branches that later merge into a single trunk (Figure 1). One of the branches reads contiguous SNPs within a genomic region involving each gene, while the other reads the covariates. The common trunk combines this information to predict family history of AD. Each sample consists of SNPs and covariates for a group of 10 individuals, with all 10 being either cases (people with family history of AD) or controls. The NN was trained on predicting whether the sample is formed by cases or controls. The rationale behind using convolutional layers in our architecture (Figure 1) was to implement "group training", which allows the NNs of GWANN to consider the group of 10 cases or

controls as a single sample, enabling them to identify similar patterns across the individuals in the group. Before passing the output of this section to the densely connected section of the model, they are passed through an “attention” block to focus on important features and ignore features without much information. The final feature vector, obtained from the densely connected portion of the NN focussing on the SNPs, is concatenated with a feature vector or encoding generated from the covariates (bottom-left branch of NN in Figure 1) and finally passed through the densely connected end layers of the NN to obtain the final prediction. The covariate encodings are obtained from the penultimate layer of the bottom-left branch. Further information about the NN architecture can be found in the Supplementary File 1.

4.4 Training the neural network

The NNs were trained to predict the status of maternal and paternal history of AD in two separate analyses. Gene locations were mapped according to the Genome Reference Consortium Human Build 37 (GRCh37/hg19). The dataset for each phenotype was split into case-control balanced training and testing sets. For maternal history, this resulted in a training set of 44,426 and a testing set of 7,840; paternal history was split into a training set of 21,556 and a testing set of 3,804. In order to implement “group training”, we separately upsampled the individuals in the training and testing sets by 10 and randomly grouped them into groups of 10. We ensured no overlap of individuals between the training and testing sets. All missing SNP values were replaced with -1 and all variables were min-max scaled prior to training. To optimise the training process, (i) we pre-trained the covariate branch (bottom-left branch in Figure 1); and (ii) saved and reused the encodings from the penultimate layer, instead of passing the covariates through the pre-trained branch during each training to obtain the same encodings.

For every gene, SNPs within the gene and in the 2500 bp flanking region (upstream and downstream of the gene) were considered. Since NNs are computationally more intensive

than linear models, we set the limit to 2500 bp as a trade-off between increased computational time and including downstream and upstream SNPs in the analysis. This also minimised the chances of overlap between genes which are very close to each other. We divided every gene into windows of maximum 50 SNPs and the final analysis was done on all windows of all genes. A different NN was trained for each window per gene in the entire genome. This resulted in having to train a total of 73,310 models for maternal history and 73,299 models for paternal history. The NN models were built and trained using PyTorch[40], a Python library built and optimised for machine learning and deep learning.

4.5 Identifying significantly associated genes

A null distribution of accuracies, A_{null}, was obtained from a set of NNs trained on the same covariate encodings along with simulated SNP data generated using the "dummy" command of PLINK 2.0[41,42]. The P value of window *i* was obtained as $1 - CDF_{null}(a_i)$, where CDF_{null} is the cumulative distribution function of the distribution fit to A_{null}, and a_i is the NN accuracy for window *i.*

4.6 Meta-analysis for parental histories of AD

After conducting separate analyses for maternal and paternal histories of AD, for each phenotype, we assigned a gene-level P value as the most significant P value amongst all windows for that gene. Following this, we used METAL[43] to perform a gene-level meta-analysis of the parental histories of AD (maternal and paternal). Since NNs do not produce beta values or standard errors, the "sample size weighted" method was used and the "weight" parameter for maternal and paternal histories were set as 7,840 and 3,804 respectively, the number of samples in the testing set for each analysis. After the meta-analysis, LDLink[44] was used to find the LD between SNPs using the "British in England and Scotland" population. A pair of genes were reported to be in LD at a given threshold if any SNPs mapped to these genes were in LD with each other.

4.7 Enrichment and post-hoc analyses

We used information from the Agora AD knowledge portal (https://agora.adknowledgeportal.org) to identify genes that have (i) significant eQTLs in the brain; (ii) change in RNA expression in post-mortem AD brains; (iii) AD target nominations; and (iv) association with AD cognitive phenotypes (Braak stage, CERAD and COGDX). STRING V11.5[45] was used to perform PPI analysis of the genes that were significantly associated after the parental meta-analysis. Pathway enrichment was performed using GSEA[46]. This enrichment was performed using the gene-level statistics for the parental meta-analysis for all analysed genes. The enrichment was performed for KEGG, Wiki and Reactome pathways present in the canonical pathways of MSigDB v7.5.1[47]. Using the same method, we also performed an enrichment analysis of DEG sets for different AD-related conditions, in different brain regions from two studies — meta-analysis of AD brain transcriptomic data[16] and GSE118553[17] —, with the exception of multiple testing corrections. Instead of using an empirical FDR, for each condition reported in the study, we applied a Bonferroni correction for the number of DEG sets (different brain regions) that were analysed. In addition to the GSEA, we performed a GO ORA of the GWANN hits using ClueGO v2.5.8[48] and a disease enrichment analysis with DisGeNET[49]. FUMA[50] was used to identify tissues enriched using the GWANN genes with P value$<1x10^{-5}$. Finally, we used TargetDB[51] to get a picture of the tractability or suitability of the novel GWANN hits for intervention by modalities such as small molecules or antibodies. The details and parameters of implementation for the different analyses can be found in the Supplementary File 1.

Data Availability

GWANN summary statistics and the summaries for all post-hoc analyses are available in the Supplementary Tables (see Supplementary File 1 for table descriptions).

STRING analysis of the PPI network of GWANN hits can be found here: https://version-11-5.string-db.org/cgi/network?networkId=bAnWnXnBtjR4

Data on brain eQTLS, RNA change in the brain, and AD pathology measures were obtained from the Agora AD knowledge portal (https://agora.adknowledgeportal.org), site version 3.0.0, and data version syn13363290-v40.

Code Availability

Code used in the analysis will be made available upon request.

Acknowledgments

We thank the UK Biobank participants and the UK Biobank team for their work in collecting, processing, and disseminating these data for analysis. This research was conducted using the UK Biobank Resource under the approved project 15181. The results published here are in part based on data obtained from Agora (https://agora.adknowledgeportal.org), a platform initially developed by the NIA-funded AMP-AD consortium that shares evidence in support of AD target discovery.

Funding: This work was supported by the Centre for Artificial Intelligence in Precision Medicines (CAIPM); Johnson and Johnson; the John Fell Foundation [grant ID 0010659]; and the Virtual Brain Cloud from European commission [grant number H2020-SC1-DTH-2018-1]. C.A. is funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Conflicts of Interest / Competing Interests

W.S. is funded by Johnson and Johnson. A.N.H. received funding from Johnson and Johnson, GlaxoSmithKline, and Ono Pharma.

Author Contributions

U.G.: Conceptualisation, Methodology, Software, Formal analysis, Visualization, Writing - Original Draft. **W.S.**: Conceptualisation, Methodology, Formal analysis, Writing - Original Draft. **L.W.**: Methodology, Investigation, Writing - Original Draft. **M.F.**: Formal analysis, Visualisation, Writing - Original Draft. **D.N.**: Investigation, Writing - Review & Editing. **B.S.U.**: Formal analysis, Writing - Review & Editing. **Q.L.**: Software. **L.S., C.A., A.A., M.A., H.C.**: Writing - Review & Editing. **C.V.D.**: Supervision, Writing - Review & Editing. **A.N.H.**: Conceptualisation, Methodology, Supervision, Writing - Review & Editing.

References

1. Serge Gauthier, Pedro Rosa-Neto, José A. Morais, & Claire Webster. *World Alzheimer Report 2021: Journey through the diagnosis of dementia*. (2021).
2. Querfurth, H. W. & LaFerla, F. M. Alzheimer's disease. *N. Engl. J. Med.* **362**, 329–344 (2010).
3. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
4. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
5. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
6. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
7. Marioni, R. E. *et al.* GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 99 (2018).
8. Alzheimer Disease Genetics Consortium (ADGC), *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
9. Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).
10. Rusk, N. Deep learning. *Nat. Methods* **13**, 35–35 (2016).
11. van Hilten, A. *et al.* GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. *Commun. Biol.* **4**, 1–9 (2021).

12. Sudlow, C. *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.* **12**, e1001779 (2015).
13. Kulminski, A. M., Loiko, E., Loika, Y. & Culminskaya, I. Pleiotropic predisposition to Alzheimer's disease and educational attainment: insights from the summary statistics analysis. *GeroScience* **44**, 265–280 (2022).
14. DeMichele-Sweet, M. A. A. *et al.* Genome-wide association identifies the first risk loci for psychosis in Alzheimer disease. *Mol. Psychiatry* **26**, 5797–5811 (2021).
15. Dumitrescu, L. *et al.* Sex differences in the genetic predictors of Alzheimer's pathology. *Brain J. Neurol.* **142**, 2581–2589 (2019).
16. Patel, H., Dobson, R. J. B. & Newhouse, S. J. A Meta-Analysis of Alzheimer's Disease Brain Transcriptomic Data. *J. Alzheimers Dis. JAD* **68**, 1635–1656 (2019).
17. Patel, H. *et al.* Transcriptomic analysis of probable asymptomatic and symptomatic alzheimer brains. *Brain. Behav. Immun.* **80**, 644–656 (2019).
18. Cruchaga, C. *et al.* GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's disease. *Neuron* **78**, 256–268 (2013).
19. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
20. Swaminathan, S. *et al.* Analysis of Copy Number Variation in Alzheimer's Disease: the NIA-LOAD/NCRAD Family Study. *Curr. Alzheimer Res.* **9**, 801–814 (2012).
21. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* 1–25 (2022) doi:10.1038/s41588-022-01024-z.
22. Zhang, L. *et al.* Roles and Mechanisms of Axon-Guidance Molecules in Alzheimer's Disease. *Mol. Neurobiol.* **58**, 3290–3307 (2021).
23. Rama, N. *et al.* Amyloid precursor protein regulates netrin-1-mediated commissural axon outgrowth. *J. Biol. Chem.* **287**, 30014–30023 (2012).

24. Sentürk, A., Pfennig, S., Weiss, A., Burk, K. & Acker-Palmer, A. Ephrin Bs are essential components of the Reelin pathway to regulate neuronal migration. *Nature* **472**, 356–360 (2011).
25. Sabatino, J. J., Pröbstel, A.-K. & Zamvil, S. S. B cells in autoimmune and neurodegenerative central nervous system diseases. *Nat. Rev. Neurosci.* **20**, 728–745 (2019).
26. Kim, K. *et al.* Therapeutic B-cell depletion reverses progression of Alzheimer's disease. *Nat. Commun.* **12**, 2185 (2021).
27. Dean, C. & Dresbach, T. Neuroligins and neurexins: linking cell adhesion, synapse formation and cognitive function. *Trends Neurosci.* **29**, 21–29 (2006).
28. Glessner, J. T. *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459**, 569–573 (2009).
29. Walsh, T. *et al.* Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. *Science* **320**, 539–543 (2008).
30. Kochunov, P. *et al.* A White Matter Connection of Schizophrenia and Alzheimer's Disease. *Schizophr. Bull.* **47**, 197–206 (2020).
31. Douaud, G. *et al.* A common brain network links development, aging, and vulnerability to disease. *Proc. Natl. Acad. Sci.* **111**, 17648–17653 (2014).
32. Cohen, P., Cross, D. & Jänne, P. A. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nat. Rev. Drug Discov.* **20**, 551–569 (2021).
33. Paris, D. *et al.* Inhibition of Alzheimer's β-Amyloid Induced Vasoactivity and Proinflammatory Response in Microglia by a cGMP-Dependent Mechanism. *Exp. Neurol.* **157**, 211–221 (1999).
34. Sha, M. C. & Callahan, C. M. The Efficacy of Pentoxifylline in the Treatment of Vascular Dementia: A Systematic Review. *Alzheimer Dis. Assoc. Disord.* **17**, 46–54 (2003).
35. Janitschke, D. *et al.* Methylxanthines and Neurodegenerative Diseases: An Update. *Nutrients* **13**, 803 (2021).

36. Hinton, G., Vinyals, O. & Dean, J. Distilling the Knowledge in a Neural Network. in *NIPS Deep Learning and Representation Learning Workshop* (2015).
37. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
38. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
39. Newby, D. *et al.* The relationship between isolated hypertension with brain volumes in UK Biobank. *Brain Behav.* **12**, e2525 (2022).
40. Paszke, A. *et al.* PyTorch: an imperative style, high-performance deep learning library. in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* 8026–8037 (Curran Associates Inc., 2019).
41. Purcell, Shaun & Chang, Christopher. PLINK 2.0.
42. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).
43. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
44. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
45. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
46. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
47. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

48. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinforma. Oxf. Engl.* **25**, 1091–1093 (2009).
49. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).
50. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
51. De Cesco, S., Davis, J. B. & Brennan, P. E. TargetDB: A target information aggregation tool and tractability predictor. *PloS One* **15**, e0232644 (2020).

Figure captions

Figure 1. NN architecture used in the GWANN method. The top-left branch generates a 1D encoding from the SNPs input (green), while the bottom-left branch does so for the covariate input (red). The right trunk merges the encodings of both branches to output whether the input belongs to cases (blue output) or controls (red output).

Figure 2. Manhattan plot of the parental meta-analysis. The threshold for genome wide significance (P value<6.82×10^{-7}) is indicated by the red dotted line. GWANN identified 21 known hits, 2 nominated AD targets and 72 novel gene hits. The black dotted line indicates the suggestive association threshold (P value<1×10^{-5}).

Figure 3. Overlap analysis of GWANN hits, traditional GWAS hits on the GWANN cohort, and hits in the recent GWAS on the EADB cohort[21]. (a) Venn diagram of 69 maternal, 29 paternal and 95 meta-analysis genes with P value<6.82×10^{-7} in each analysis. (b) Analysis of the intersection of hit genes between the methods on the x axis and y axis. The number of genes in the sets and intersections are specified within the brackets and the intensity of each block (darker is more significant) represents the significance of the size of the intersection set, given the 2 individual sets. Blocks in grey suggest an insignificant number of genes in the intersection. The EADB set is without the *APOE* locus, and hence the intersection with the traditional UKB GWAS (using the same data set as used for GWANN) results in a null set.

Figure 4. PPI network of GWANN hits from STRING. Red - old hits/suggestive hits, orange - nominated hits, blue - novel hits. The grey solid lines indicate a direct connection between two genes and the green dotted lines indicate an indirect connection through a known hit. In each of the networks, the node with the highest degree (most interactions) is

the bottommost node followed by decreasing degree in the anti-clockwise direction. (a) PPI network containing the 46 (inclusive of old, nominated, and novel hits) out of the 95 hits that had connections with each other. (b) The same network as in (a) after removing all known hits.

Figure 5. Go term, disease, and tissue enrichment of the GWANN hits. (a) GO term over representation analysis for GWANN meta-analysis hits using ClueGO v2.5.8. For network visualisation purposes, GO child terms spanning more than two were grouped and collapsed into a node (white nodes). V-shaped gene nodes correspond to either antisense, '*BDNF-AS*' as opposed to *BDNF*; or flanking sequence regions, between '*PDCD10*' and '*SERPINI1*'. GO node size is proportional to the number of genes enriched in the group. (b) Pie chart showing the percentage of GWANN hits in each GO term. (c) Dotplot of the disease enrichment analysis with DisGeNET, with gene ratio (x-axis, #input genes / #genes in disease term); disease terms from a multitude of disease-gene databases (y-axis); false discovery rate (FDR) for each disease enriched term; and number of genes (#genes found in the disease term). (d) Tissue specificity plot (up and down regulated) for GTEx v8 54 tissue types using GWANN genes with P value<$1x10^{-5}$ in the paternal meta-analysis. Significantly enriched tissues (P_{bon}<0.05) are highlighted in red.

Tables

Table 1. GWANN hit genes after the meta-analysis.

HGNC Symbol	CHR	P_{m+p}	P_m	P_p	HGNC Symbol	CHR	P_{m+p}	P_m	P_p
DAB1 [k]	1	6.50E-08	9.00E-06	2.00E-03	PITPNM2 [k]	12	7.50E-08	6.10E-10	6.00E-01
NRXN1 [k]	2	3.40E-09	4.20E-06	1.90E-04	NRXN3 [k]	14	2.80E-08	2.60E-05	2.30E-04
LRP1B [k]	2	3.40E-08	1.90E-05	4.50E-04	APH1B [k]	15	3.80E-11	4.40E-09	1.70E-03
ARHGAP15 [k]	2	7.80E-08	2.60E-05	7.60E-04	WWOX [k]	16	1.30E-10	6.90E-08	4.50E-04
RBMS3 [k]	3	9.30E-08	4.90E-04	1.40E-05	DCC [k]	18	2.10E-08	2.90E-06	2.00E-03
CADM2 [k]	3	2.30E-07	1.90E-05	3.60E-03	APOE [k]	19	9.60E-190	1.00E-65	1.80E-158
EFNA5 [k]	5	2.60E-08	1.80E-08	9.70E-02	TOMM40 [k]	19	9.60E-170	1.00E-65	2.40E-127
EPHA1-AS1 [k]	7	2.70E-07	2.20E-02	1.20E-08	APOC1 [k]	19	9.40E-153	1.00E-65	2.50E-102
EXOC4 [k]	7	6.50E-08	8.80E-07	1.60E-02	BCAM [k]	19	7.50E-75	1.00E-64	2.20E-14
GLIS3 [k]	9	4.10E-20	6.90E-08	8.40E-17	GEMIN7 [k]	19	1.60E-21	2.60E-28	4.00E-01
PAX5 [k]	9	2.80E-08	1.30E-06	5.60E-03	EXOC3L2 [k]	19	8.90E-08	3.50E-08	1.50E-01
ANK3 [k]	10	3.60E-10	2.80E-08	2.70E-03	PPP1R37 [k]	19	1.90E-08	6.90E-08	3.70E-02
LUZP2 [k]	11	1.90E-10	1.30E-06	2.70E-05	MACROD2 [k]	20	4.80E-11	5.30E-05	1.20E-08
FGGY	1	2.60E-10	2.20E-10	5.20E-02	VIPR2	7	2.80E-09	8.80E-04	1.90E-08
NFIA	1	5.40E-13	1.40E-14	1.20E-01	CRISPLD1	8	4.10E-07	2.80E-08	3.70E-01
PIGK	1	6.40E-08	6.90E-08	8.60E-02	EXT1	8	3.10E-08	2.00E-06	4.30E-03
FSHR	2	6.80E-10	1.40E-02	3.50E-13	PEBP4	8	4.40E-07	3.20E-02	8.30E-09
LYPD6B	2	5.80E-07	2.90E-06	4.20E-02	SGCZ	8	2.80E-08	5.90E-07	1.10E-02
MYT1L	2	1.60E-11	1.10E-08	3.20E-04	NTRK2	9	1.40E-07	2.90E-06	1.30E-02
PARD3B	2	1.90E-10	2.50E-07	1.90E-04	CTNNA3	10	1.20E-07	5.30E-05	5.50E-04
TANC1	2	2.80E-08	1.30E-06	5.60E-03	HPSE2	10	3.00E-11	1.50E-11	5.20E-02
ATP1B3	3	5.60E-09	2.70E-09	9.70E-02	WDR11-AS1	10	5.80E-07	5.90E-07	1.20E-01
ATP2C1	3	1.20E-07	5.90E-07	3.70E-02	BDNF	11	3.00E-08	6.10E-06	1.40E-03
CHCHD6	3	3.20E-07	2.00E-03	6.50E-06	CNTN5	11	3.80E-07	4.90E-04	1.00E-04
CNTN4	3	2.70E-11	2.80E-08	2.30E-04	ELP4	11	1.80E-08	2.60E-05	1.30E-04
EPHA6	3	1.90E-07	6.20E-06	8.60E-03	LRRC4C	11	4.90E-07	1.30E-05	1.10E-02
FLNB	3	6.50E-08	9.00E-06	2.00E-03	USP47	11	2.10E-08	4.40E-03	1.20E-08
IQSEC1	3	4.70E-11	4.40E-09	2.00E-03	MGAT4C	12	8.20E-09	1.10E-08	5.90E-02
ROBO1	3	2.10E-08	1.80E-08	8.60E-02	C1QTNF9	13	1.30E-07	5.60E-08	1.50E-01
SERPINI1	3	1.70E-07	7.40E-05	5.50E-04	MTUS2	13	8.90E-12	1.70E-09	1.00E-03
STAC	3	4.90E-07	2.60E-05	5.60E-03	NALCN	13	2.50E-07	7.10E-03	2.40E-07
UBE2E2	3	5.40E-08	4.20E-06	3.60E-03	SMAD9	13	1.90E-12	3.40E-16	5.40E-01
ARHGAP24	4	3.10E-08	4.40E-09	2.10E-01	GPR137C	14	4.90E-07	6.90E-08	2.90E-01
SORCS2	4	4.50E-08	7.40E-05	1.00E-04	SAMD4A	14	5.20E-08	1.80E-08	1.50E-01
ARHGEF28	5	6.60E-11	6.90E-08	2.30E-04	CORO2B	15	1.30E-08	2.50E-07	1.10E-02
COL23A1	5	2.40E-08	7.10E-03	3.50E-09	UBE3A	15	1.30E-07	1.00E-09	6.30E-01
TENM2	5	1.80E-12	6.20E-06	5.00E-09	MYH11	16	6.30E-07	2.60E-05	7.30E-03
CPNE5	6	2.60E-08	1.80E-08	9.70E-02	TMEM170A	16	1.60E-07	6.90E-08	1.50E-01

KIF13A	6	1.60E-08	7.00E-09	1.20E-01	ARSG	17	1.90E-15	1.30E-18	2.10E-01
RNF182	6	3.00E-07	4.90E-04	7.60E-05	ASIC2	17	1.60E-08	1.80E-08	7.20E-02
BAIAP2L1	7	1.90E-07	1.10E-07	1.40E-01	BAHCC1	17	1.80E-08	2.80E-08	5.90E-02
CARD11	7	7.70E-08	1.30E-05	1.70E-03	LMAN1	18	1.60E-08	1.10E-07	2.40E-02
IMMP2L	7	1.60E-07	2.00E-04	1.30E-04	ICAM3	19	6.40E-08	6.90E-08	8.60E-02
MAGI2	7	8.90E-12	1.70E-09	1.00E-03	SNRPB2	20	2.50E-12	4.80E-14	1.50E-01
PDE1C	7	6.60E-11	6.10E-10	1.10E-02	C2CD2	21	7.70E-14	4.40E-09	3.30E-06
PTPRN2	7	5.60E-09	1.90E-05	5.10E-05	LINC00113	21	3.70E-09	7.70E-11	3.30E-01
TYW1B	7	5.90E-08	2.90E-06	5.60E-03	BCR	22	8.20E-09	1.70E-09	1.50E-01

P_{m+p} represents the METAL meta-analysis P value, P_m the GWANN maternal analysis P value and P_p the GWANN paternal analysis P value.

k – Known gene, associated with AD or AD-related traits in previous GWAS

Table 2. GWANN hit genes that had significant association with AD pathology measures.

HGNC Symbol	Brain eQTL	RNA change in brain	Braak (adj_p)	CERAD (adj_p)	COGDX (adj_p)
APOC1	**TRUE**	**TRUE**	**8.85×10^{-03}**	**3.29×10^{-02}**	5.87×10^{-02}
ARSG	**TRUE**	**TRUE**	7.55×10^{-01}	6.38×10^{-01}	**2.24×10^{-02}**
MAGI2	**TRUE**	**TRUE**	3.91×10^{-01}	5.70×10^{-01}	**3.79×10^{-02}**
PDE1C	**TRUE**	FALSE	2.35×10^{-01}	**6.99×10^{-03}**	9.28×10^{-01}
ATP1B3	**TRUE**	**TRUE**	3.62×10^{-01}	2.35×10^{-01}	**1.09×10^{-02}**
CPNE5	**TRUE**	**TRUE**	7.24×10^{-02}	2.20×10^{-01}	**3.39×10^{-02}**
NRXN3	**TRUE**	**TRUE**	6.92×10^{-01}	7.34×10^{-01}	**2.32×10^{-02}**
SAMD4A	**TRUE**	**TRUE**	**2.61×10^{-05}**	**3.11×10^{-04}**	**2.60×10^{-05}**
PITPNM2	**TRUE**	FALSE	**1.69×10^{-02}**	7.54×10^{-02}	1.60×10^{-01}
RNF182	**TRUE**	FALSE	**4.58×10^{-02}**	**1.32×10^{-03}**	1.85×10^{-01}
LYPD6B	**TRUE**	TRUE	6.58×10^{-01}	**4.41×10^{-02}**	8.45×10^{-01}

The pathology measures considered were Braak stage, CERAD and COGDX. All GWANN hits that had significant association with at least one of these measures have been reported. The data presented in this table was obtained from the Agora AD knowledge portal (https://agora.adknowledgeportal.org/).

Table 3. **Enrichment of DEGs identified in two transcriptomic studies on AD brains.**

a	AD	Other brain disorders	Only AD	AD and other brain disorders
Cerebellum	**$1.37x10^{-3}$***	**$1.04x10^{-3}$***	**$5.49x10^{-3}$***	**$4.13x10^{-3}$***
Parietal lobe	**$8.77x10^{-7}$***	**$7.31x10^{-11}$***	**$5.82x10^{-4}$***	**$4.67x10^{-5}$***
Frontal lobe	**$6.38x10^{-3}$***	**$6.46x10^{-5}$***	**$1.10x10^{-2}$***	$2.01x10^{-2}$
Temporal lobe	$5.82x10^{-2}$	$4.06x10^{-1}$	$5.82x10^{-2}$	-

b	AD vs control	Asymptomatic AD vs AD	Asymptomatic AD vs Controls
Cerebellum	$6.11x10^{-1}$	$8.26x10^{-1}$	$1.46x10^{-1}$
Entorhinal cortex	**$3.11x10^{-5}$***	**$9.39x10^{-5}$***	$5.53x10^{-1}$
Frontal cortex	$4.34x10^{-2}$	$3.89x10^{-2}$	$1.55x10^{-1}$
Temporal cortex	**$1.49x10^{-3}$***	$2.24x10^{-2}$	$5.93x10^{-2}$

The enrichments were performed using the GWANN meta-analysis summary statistics. (a) Enrichment P values of DEGs in four different brain regions for AD and other mental disorders analysed by Patel et al.[16] (b) Enrichment P values of DEGs in four different brain regions for the analysis between AD, asymptomatic AD and controls by Patel et al.[17]

* P value<$1.25x10^{-2}$

Table 4. Disease enrichment of the GWANN hit genes using the DisGeNET database.

Disease	FDR	GWANN hit count	Disease	FDR	GWANN hit count
Alzheimer's Disease	2.42×10^{-05}	38	Adolescent Idiopathic Scoliosis	1.05×10^{-07}	19
Schizophrenia	1.38×10^{-06}	37	Bipolar Disorder	3.54×10^{-04}	19
Mathematical Ability	6.82×10^{-14}	29	Systolic Pressure	2.01×10^{-05}	18
Obesity	2.36×10^{-03}	29	Autism Spectrum Disorders	3.54×10^{-04}	18
Body Mass Index	2.55×10^{-11}	28	Major Depressive Disorder	4.49×10^{-03}	17
Leukaemia, Myelocytic, Acute	3.70×10^{-02}	27	Impaired Cognition	4.81×10^{-02}	17
Intellectual Disability	7.02×10^{-04}	26	Abnormal Behaviour	6.18×10^{-03}	14
Autistic Disorder	6.67×10^{-07}	23	Familial (FPAH)	2.26×10^{-02}	14
Seizures	8.28×10^{-03}	23	Blood Protein Measurement	3.78×10^{-02}	14
Intelligence	1.58×10^{-09}	21	Attention Deficit Hyperactivity Disorder	9.41×10^{-03}	13
Neurodevelopmental Disorders	5.72×10^{-10}	20	Anxiety	4.30×10^{-02}	13
Global Developmental Delay	1.56×10^{-02}	20	Developmental Delay (Disorder)	1.74×10^{-03}	12
Body Height	2.26×10^{-02}	20	Alzheimer Disease, Late Onset	3.28×10^{-04}	11
Smoking	2.94×10^{-11}	19	Mental Disorders	4.24×10^{-02}	11
Scoliosis, Isolated, Susceptibility To, 3	1.43×10^{-08}	19	Dementia	4.93×10^{-02}	11

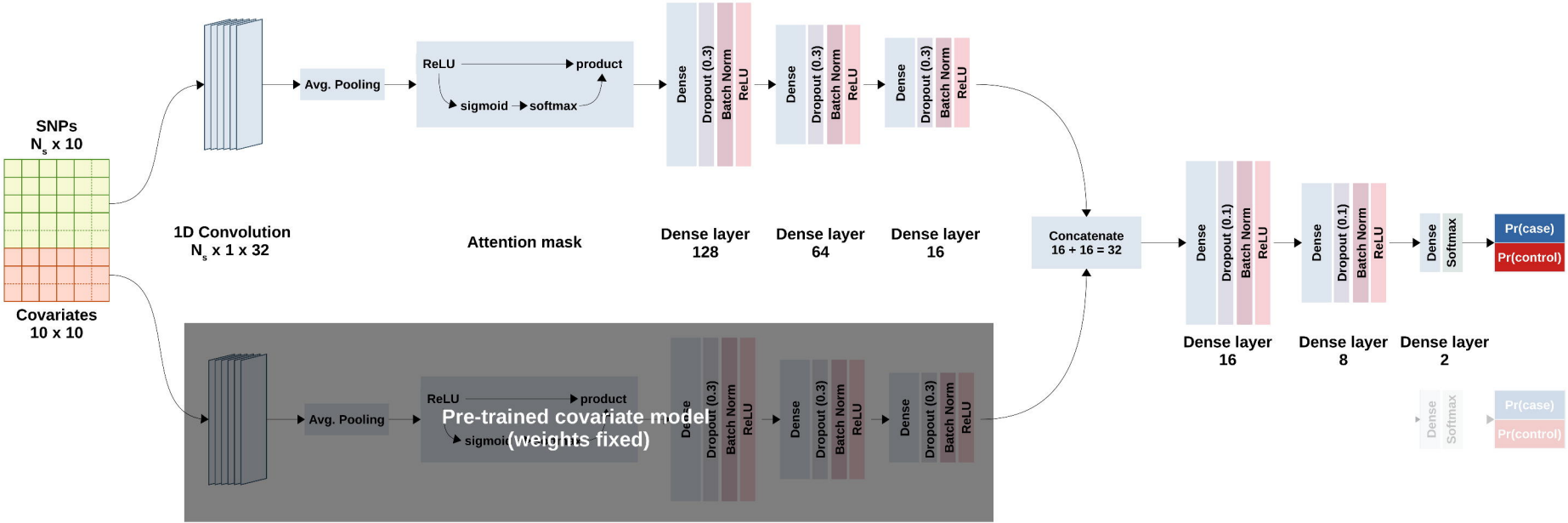
Gene sets belonging to different diseases and traits that were significantly enriched after applying a false discovery rate (FDR) correction. Column “GWANN hit count” indicates the number of GWANN hit genes that were part of the disease gene sets.

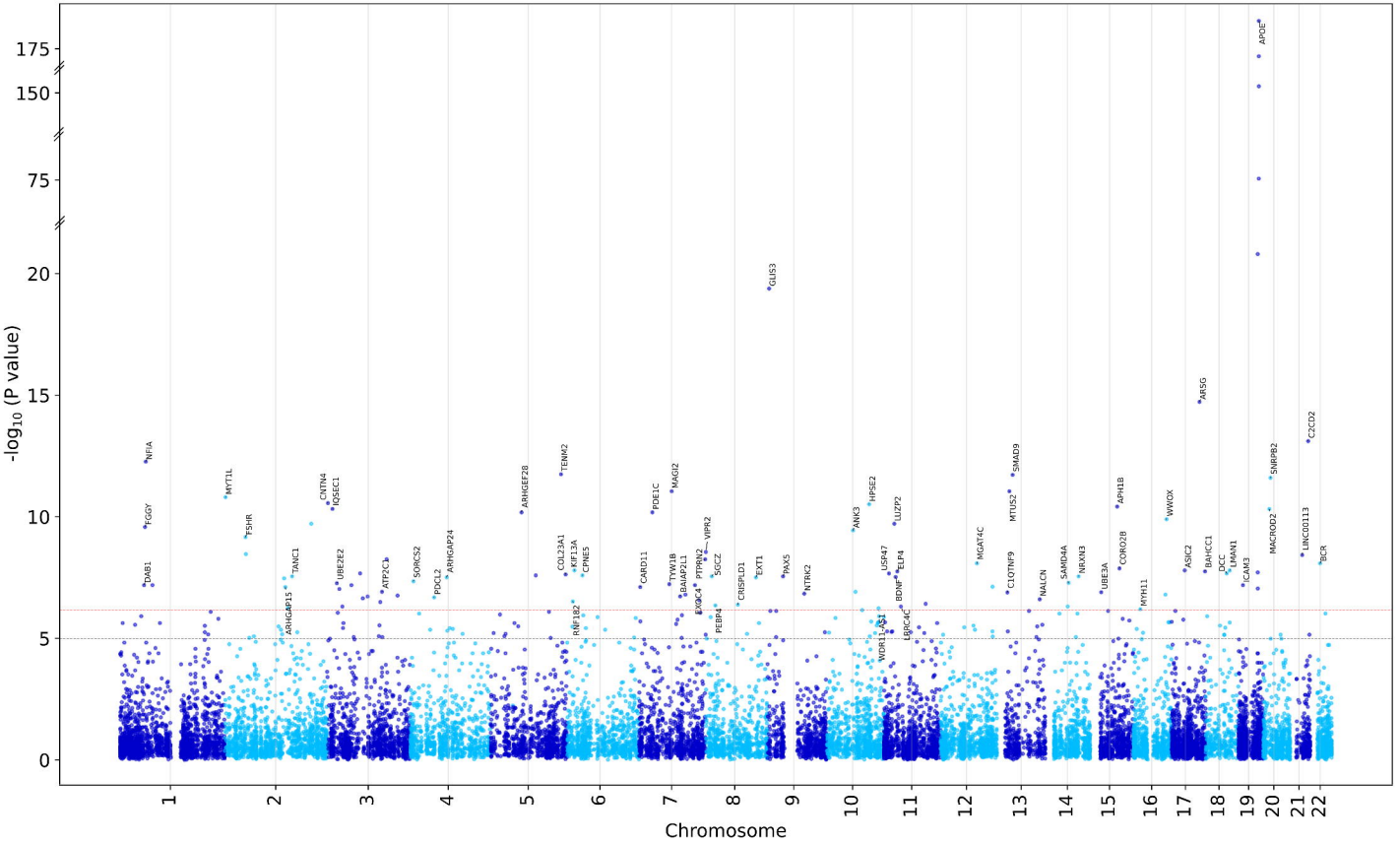
Table 5. Novel GWANN hits with known associated drugs and the main category of diseases they are used for.

HGNC symbol	Drug Name	Type	Associated disease
BCR	Bosutinib	Small molecule	Cancer
	Dasatinib	Small molecule	Cancer
	Imatinib	Small molecule	Cancer
	Imatinib mesylate	Small molecule	Cancer
	Nilotinib	Small molecule	Cancer
	Nilotinib hydrochloride monohydrate	Small molecule	Cancer
	Ponatinib	Small molecule	Cancer
	Ponatinib hydrochloride	Small molecule	Cancer
	Radotinib	Small molecule	Cancer
EPHA6	Vandetanib	Small molecule	Cancer
FSHR	Corifollitropin alfa	Protein	Infertility
			Obesity
			Hypogonadotropic hypogonadism
	Follitropin alfa	Protein	Infertility
			Polycystic ovary syndrome
	Menotropins	Protein	Infertility
			Hypogonadotropic hypogonadism
			Polycystic ovary syndrome
			Kallmann syndrome
	Urofollitropin	Unknown	Infertility
NTRK2	Altiratinib	Small molecule	Cancer
	AZD-6918	Small molecule	Cancer
	AZD-7451	Small molecule	Glioblastoma multiforme
	Cenegermin	Protein	Cancer
	Entrectinib	Small molecule	Cancer
			Neoplasm
	Larotrectinib	Small molecule	Cancer
			Neoplasm
	Larotrectinib sulfate	Small molecule	Leukaemia
	Lestaurtinib	Small molecule	Giant cell tumour
	PLX-7486	Small molecule	Neoplasm

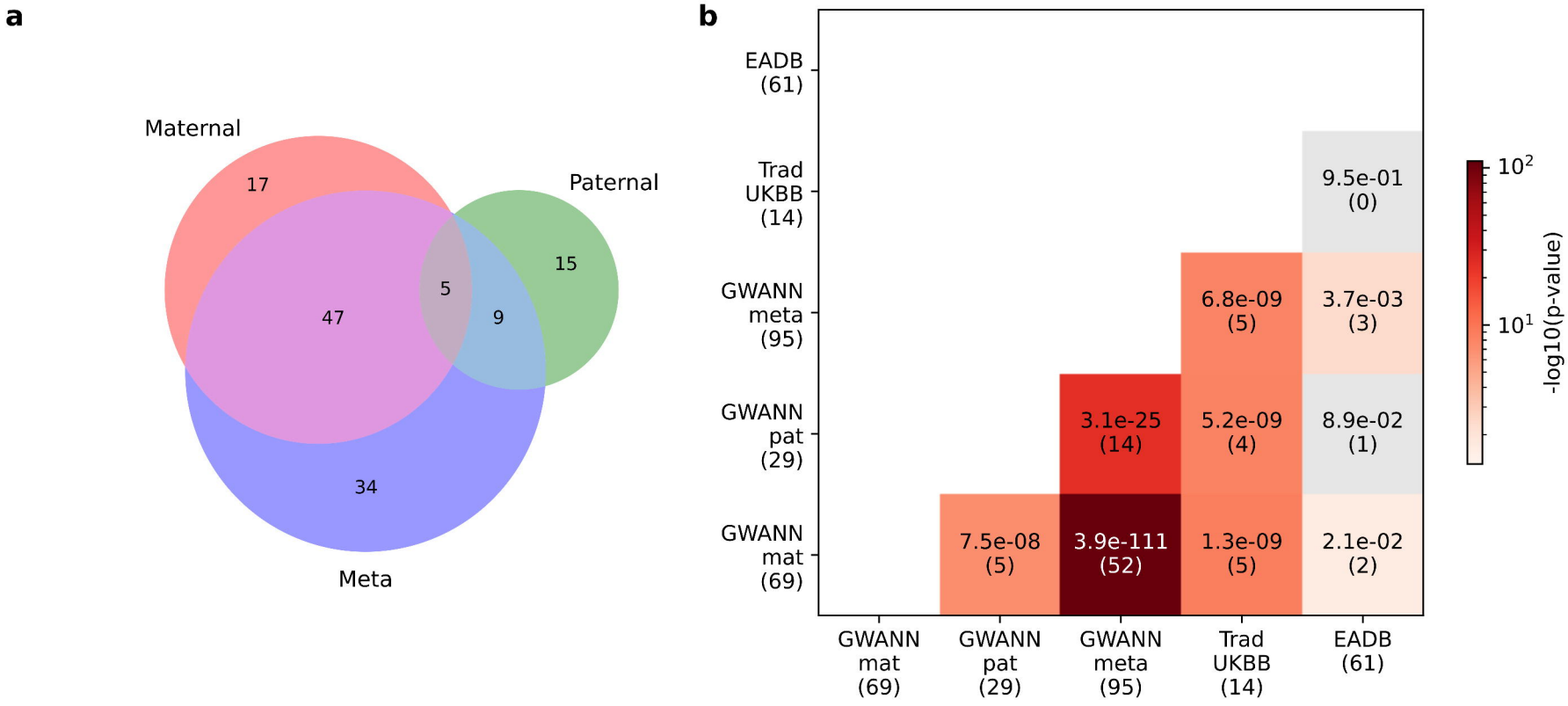
HGNC symbol	Drug Name	Type	Associated disease
	Selitrectinib	Small molecule	Keratitis, eye disease
PDE1C	Dipyridamole	Small molecule	Stroke
			Heart disease
			Bipolar disorder
			Diabetes mellitus
			COVID-19
	Pentoxifylline	Small molecule	Diabetes mellitus
			Brain disease
			Hepatic disease

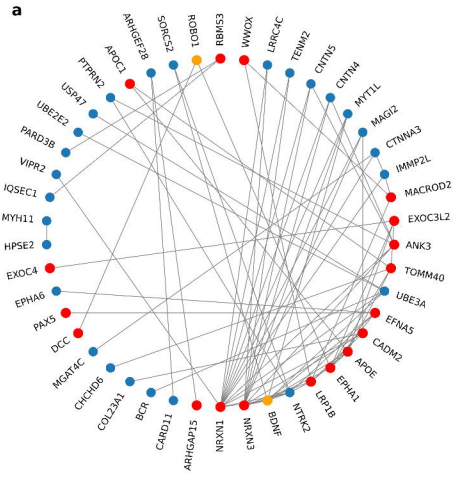
TargetDB was used to identify the hits with known associated drugs, and the specific list of drugs and diseases were obtained from OpenTargets.

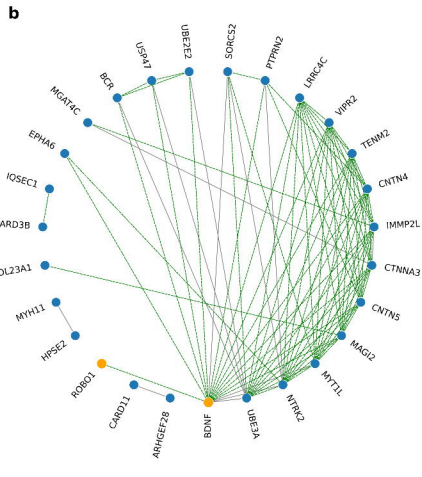

SNPs
N_s x 10
Covariates
10 x 10
1D Convolution
N_s x 1 x 32
Avg. Pooling
ReLU
product
sigmoid
softmax
Attention mask
Dense
Dropout (0.3)
Batch Norm
ReLU
Dense layer
128
Dense layer
64
Dense layer
16
Concatenate
16 + 16 = 32
Dropout (0.1)
Dense layer
16
Dense layer
8
Dense layer
2
Softmax
Pr(case)
Pr(control)
Pre-trained covariate model
(weights fixed)

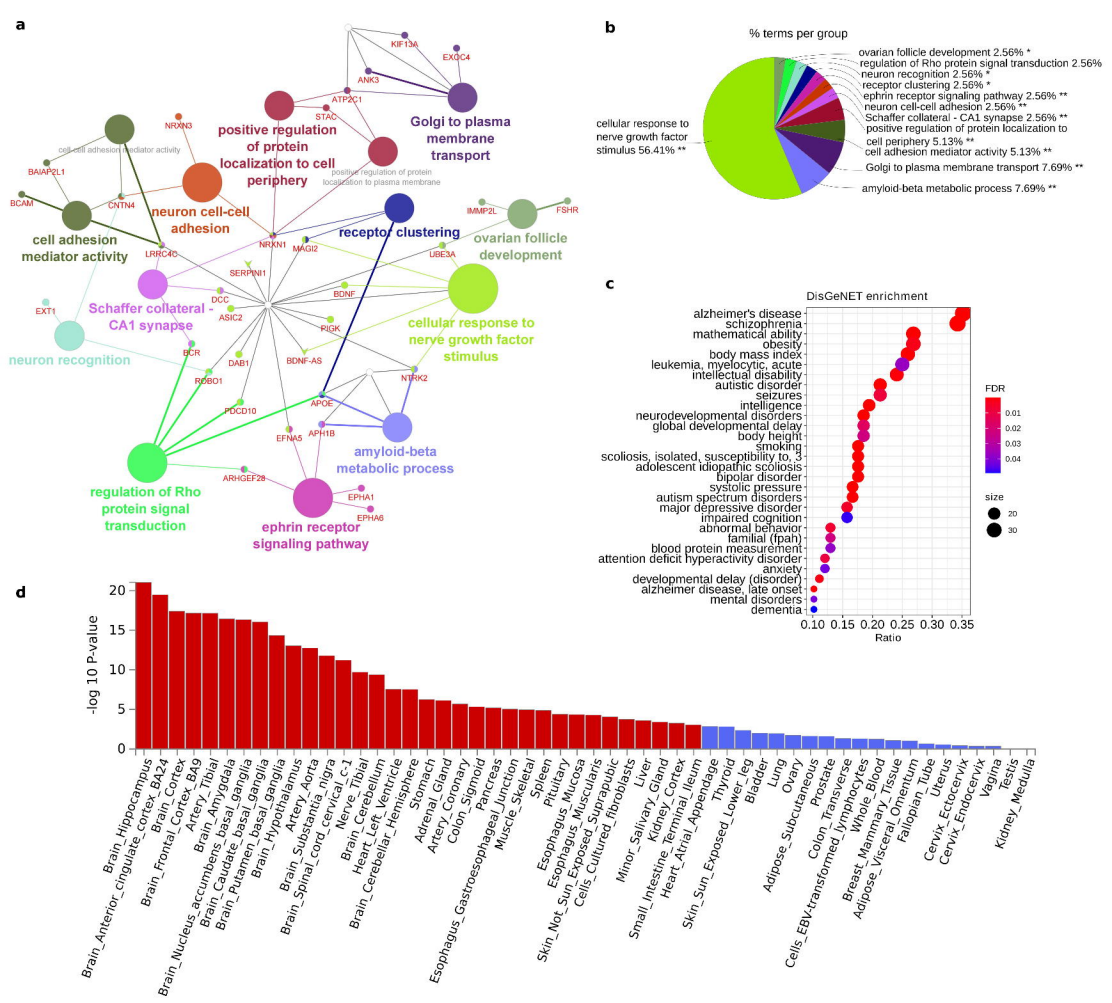


-log10 (P value)
Chromosome
NFIA
FGGY
DAB1
MYT1L
FSHR
ARHGAP15
TANC1
CNTN4
IQSEC1
UBE2E2
ATP2C1
SORCS2
PDCL2
ARHGAP24
ARHGEF28
COL23A1
KIF13A
CPNE5
RNF182
TENM2
CARD11
PDE1C
TYW1B
BAIAP2L1
PTPRN2
VIPR2
EXOC4
SGCZ
PEBP4
MAGI2
CRISPLD1
EXT1
GLIS3
PAX5
NTRK2
ANK3
HPSE2
USP47
ELP4
BDNF
WDR11-AS1
LBRC4C
LUZP2
MGAT4C
C1QTNF9
MTUS2
SMAD9
NALCN
SAMD4A
NRXN3
UBE3A
COROZB
APH1B
MYH11
WWOX
ASIC2
BAHCC1
DCC
LMAN1
ICAM3
ARSG
MACROD2
SNRPB2
LINC00113
BCR
C2CD2
APOE


a
Maternal
Paternal
Meta
17
15
5
47
9
34
b
EADB (61)
Trad UKBB (14)
GWANN meta (95)
GWANN pat (29)
GWANN mat (69)
9.5e-01 (0)
6.8e-09 (5)
3.7e-03 (3)
3.1e-25 (14)
5.2e-09 (4)
8.9e-02 (1)
7.5e-08 (5)
3.9e-111 (52)
1.3e-09 (5)
2.1e-02 (2)
GWANN mat (69)
GWANN pat (29)
GWANN meta (95)
Trad UKBB (14)
EADB (61)
10^2
10^1
-log10(p-value)


a
ARHGEF28
SORCS2
ROBO1
RBM53
WWOX
LRRC4C
TENM2
CNTN5
CNTN4
MYT1L
MAGI2
CTNNA3
IMMP2L
MACROD2
EXOC3L2
ANK3
TOMM40
UBE3A
EFNA5
CADM2
APOE
EPHA1
LRP1B
NTRK2
BDNF
NRXN3
NRXN1
ARHGAP15
CARD11
BCR
COL23A1
CHCHD6
MGAT4C
DCC
PAX5
EPHA6
EXOC4
HPSE2
MYH11
IQSEC1
VIPR2
PARD3B
UBE2E2
USP47
PTPRN2
APOC1


b
USP47
UBE2E2
SORCS2
PTPRN2
LRRC4C
VIPR2
TENM2
CNTN4
IMMP2L
CTNNA3
CNTN5
MAGI2
MYT1L
NTRK2
UBE3A
BDNF
ARHGEF28
CARD11
ROBO1
HPSE2
MYH11
COL23A1
PARD3B
IQSEC1
EPHA6
MGAT4C
BCR


a
KIF13A
EXOC4
ANK3
ATP2C1
STAC
Golgi to plasma membrane transport
positive regulation of protein localization to cell periphery
positive regulation of protein localization to plasma membrane
NRXN3
cell-cell adhesion mediator activity
BAIAP2L1
BCAM
CNTN4
neuron cell-cell adhesion
cell adhesion mediator activity
IMMP2L
FSHR
receptor clustering
ovarian follicle development
NRXN1
MAGI2
UBE3A
LRRC4C
SERPINI1
DCC
ASIC2
BDNF
Schaffer collateral - CA1 synapse
EXT1
PIGK
cellular response to nerve growth factor stimulus
neuron recognition
BCR
DAB1
BDNF-AS
ROBO1
NTRK2
PDCD10
APOE
EFNA5
APH1B
amyloid-beta metabolic process
ARHGEF28
regulation of Rho protein signal transduction
EPHA1
EPHA6
ephrin receptor signaling pathway
b
% terms per group
ovarian follicle development 2.56% *
regulation of Rho protein signal transduction 2.56%
neuron recognition 2.56% *
receptor clustering 2.56% *
ephrin receptor signaling pathway 2.56% **
neuron cell-cell adhesion 2.56% **
Schaffer collateral - CA1 synapse 2.56% **
positive regulation of protein localization to cell periphery 5.13% **
cell adhesion mediator activity 5.13% **
Golgi to plasma membrane transport 7.69% **
amyloid-beta metabolic process 7.69% **
cellular response to nerve growth factor stimulus 56.41% **
c
DisGeNET enrichment
alzheimer's disease
schizophrenia
mathematical ability
obesity
body mass index
leukemia, myelocytic, acute
intellectual disability
autistic disorder
seizures
intelligence
neurodevelopmental disorders
global developmental delay
body height
smoking
scoliosis, isolated, susceptibility to, 3
adolescent idiopathic scoliosis
bipolar disorder
systolic pressure
autism spectrum disorders
major depressive disorder
impaired cognition
abnormal behavior
familial (fpah)
blood protein measurement
attention deficit hyperactivity disorder
anxiety
developmental delay (disorder)
alzheimer disease, late onset
mental disorders
dementia
0.10 0.15 0.20 0.25 0.30 0.35
Ratio
FDR
0.01
0.02
0.03
0.04
size
20
30
d
-log 10 P-value
0
5
10
15
20
Brain_Hippocampus
Brain_Anterior_cingulate_cortex_BA24
Brain_Cortex
Brain_Frontal_Cortex_BA9
Artery_Tibial
Brain_Amygdala
Brain_Nucleus_accumbens_basal_ganglia
Brain_Caudate_basal_ganglia
Brain_Putamen_basal_ganglia
Brain_Hypothalamus
Artery_Aorta
Brain_Substantia_nigra
Brain_Spinal_cord_cervical_c-1
Nerve_Tibial
Brain_Cerebellum
Heart_Left_Ventricle
Brain_Cerebellar_Hemisphere
Stomach
Adrenal_Gland
Artery_Coronary
Colon_Sigmoid
Pancreas
Muscle_Skeletal
Spleen
Pituitary
Esophagus_Mucosa
Esophagus_Muscularis
Esophagus_Gastroesophageal_Junction
Cells_Cultured_fibroblasts
Skin_Not_Sun_Exposed_Suprapubic
Liver
Minor_Salivary_Gland
Kidney_Cortex
Small_Intestine_Terminal_Ileum
Heart_Atrial_Appendage
Skin_Sun_Exposed_Lower_Leg
Thyroid
Bladder
Lung
Ovary
Adipose_Subcutaneous
Prostate
Colon_Transverse
Cells_EBV-transformed_lymphocytes
Whole_Blood
Breast_Mammary_Tissue
Adipose_Visceral_Omentum
Fallopian_Tube
Uterus
Cervix_Ectocervix
Cervix_Endocervix
Vagina
Testis
Kidney_Medulla