medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 [1]	danielle.newby@psych.ox.ac.uk
Brittany Ulm [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, 3]	cassandra.adams@cmd.ox.ac.uk
Ashwag Albukhari [2, 4]	aalbukhari@kau.edu.sa
Majid Almansouri [2, 5]	malmansouri@kau.edu.sa
Hani Choudhry [2, 4]	hchoudhry@kau.edu.sa
Cornelia van Duijn [2, 6, 7]	cornelia.vanduijn@ndph.ox.ac.uk
Alejo Nevado-Holgado [1, 2, 6]	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 Medicines Discovery, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7BN, United Kingdom.
[4] Biochemistry Department, Faculty of Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia.
[5] Clinical Biochemistry Department, Faculty of Medicine, King Abdulaziz University, Jeddah 21589, Saudi Arabia.
[6] Big Data Institute, University of Oxford, Oxford OX3 7LF, United Kingdom.
[7] Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, United Kingdom.
* Corresponding Author

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Augmenting traditional genome wide association studies (GWAS) with advanced machine learning algorithms can allow the detection of novel signals in available cohorts. 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 25 known AD genes, 2 target nominations and 68 potentially 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 illustrates their potential to complement existing algorithms and methods, and enable the discovery of novel and tractable targets for AD.

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

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

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 disease pathways associated with AD[3]. However, some limitations of the current approaches at GWAS may be hindering their ability to uncover the complex genetic landscape of AD without extending the sample size. Despite the number of SNPs identified until today, they still only explain a small fraction of the heritability of the disease[4]; the hits often appear to be unrelated to each other[5,6]; they have limited biological relevance to the disease[7]; and it is often challenging to ascertain the genes and the biological mechanisms underlying each SNP association[4–7]. Along with the modern availability of large datasets[8–10], an enhancement that can complement current GWAS methods is the introduction of machine learning analysis methods that can unveil more complex patterns in genomic data that would otherwise have been missed by the traditional linear models.

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[11], especially where large amounts of data is available. Specifically, in the task of finding disease associated loci, NNs have recently been employed and tested on a list of complex traits and diseases such as eye colour and schizophrenia[12]. Given the recent success of NNs in a myriad of fields including genomics, our aim was to develop NNs specialised to perform a gene-level GWAS using SNP data available in the UK Biobank (UKBB)[13]. Our method is gene-based and considers groups of SNPs within and around a gene to establish the association of the gene with the phenotype of interest. In this work we demonstrate the application of our new GWANN method to identify associations

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

with familial history of AD, a proxy for AD[8], present the genetic associations found by the method, and systematically validate the results against brain eQTLs, AD phenotype associations, biological pathways, disease associations and differentially expressed gene sets in the AD brain.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

2 Results

2.1 Identification of novel genes related to AD using GWANN

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). All these were significant after multiple-comparison correction with a Bonferroni threshold of *P value*$<6.82\times10^{-7}$ (Figure 2, Supplementary Table 2). Table 1 lists the genes with their P values. Supplementary Table 2 additionally lists the top three input features that were most important to the NNs of GWANN in the maternal and paternal history datasets. In terms of the NN architecture, each of these features can be a SNP or a covariate encoding. Supplementary Table 3 contains the genes in linkage disequilibrium (LD) — thresholds of $r^2 \geq 0.1$ and $r^2 \geq 0.25$ — based on the top three maternal and paternal SNPs in the windows for the genes. A more conservative set of hits considering only those genes that passed P value thresholds of 0.05, 10^{-2}, 10^{-3}, 10^{-4} and 10^{-5} can be seen in Supplementary Figure 1. A standard SNP-level GWAS performed using PLINK 2.0 with a logistic regression model, on the same set of SNPs and individuals as used in GWANN, identified 13 genes for maternal history, four genes for paternal history and 14 genes in the meta-analysis. 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 and AD diagnosis listed in the GWAS catalogue[3], there was an overlap of 25 genes with the GWANN hits. Twelve of these have been previously associated with AD (*ANK3, APH1B, GLIS3, EPHA1, WWOX* in addition to the APOE locus). Seven others are listed in the GWAS catalogue as suggestive associations (*P value*$<1\times10^{-5}$) with AD (*EXOC4, NRXN1, NRXN3, PAX5, CADM2, LUZP2* and *RBMS3*). In a recent analysis on the pleiotropic predisposition to AD and educational attainment[14], *DCC* and *CADM2* were identified as associations, and *EFNA5* and *LRP1B* as suggestive associations.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

DAB1 has previously been associated with psychosis in AD[15], and *MACROD2* with neurofibrillary tangles[16].

EFNA5, *NRXN3*, *BDNF* and *ROBO1*, have been nominated as potential targets for AD. Among the other novel genes identified by our method, *PITPNM2, RNF182* and *SAMD4A* have been reported to be significantly associated with Braak; *RNF182* and *SAMD4A* are associated with CERAD; and *CPNE5, SAMD4A, ATP1B3, MAGI2* and *ARSG* associated with COGDX, three measures of AD pathology or cognitive phenotypes (Table 2). A total of 69 genes were significantly enriched with eQTLs in brain regions and 51 genes show differential RNA expression in post-mortem AD brains (Supplementary Table 4). There is evidence in the GWAS catalogue for 42 genes which have previously been associated with diseases in the category of “nervous system disease” (EFO:0000618).

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 (*P value=3.33x10*$^{-16}$*,* Figure 4a). Given a network of 95 possible nodes, the expected number of connections by chance would be 23, but the network of the GWANN hits show 72 connections, thereby being significantly enriched. On retaining only the novel GWANN genes, the PPI network remains enriched (*.0285)*. The reason for the drop in significance is possibly due to the absence of indirect interactions through the old hits, indicated by the green dotted lines in Figure 4b.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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[17], we verified whether the results of our NN analyses show significant representation 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 disorders). For each set, the DEGs in the cerebellum and parietal lobe, and for sets i, ii and iii, the frontal lobe was enriched at a significance level of 1.25×10^{-2} (=*.05/4*). For the frontal 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 compared gene expression between controls, asymptomatic AD cases and AD cases (GSE118553)[18]. The studied brain regions were the cerebellum, entorhinal cortex, frontal cortex and temporal cortex. None of the regions showed enrichment in the case of asymptomatic AD vs controls. DEGs for asymptomatic AD vs AD and AD vs controls in the entorhinal cortex showed enrichment, and DEGs for AD vs controls in the temporal cortex were enriched at a significance level of 1.25×10^{-2} (=*.05/4*). Supplementary Figures 2 and 3, respectively, contain the cumulative distribution plots used in the enrichment of each dataset.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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

The gene set enrichment analysis (GSEA) resulted in enrichment of the pathways for amyotrophic lateral sclerosis (*P value=3.76x10*$^{-4}$) and Alzheimer's disease (*2.98x10*$^{-3}$) amongst the set of KEGG pathways. Although both of these are disease-specific pathways, all KEGG pathways (disease specific and non-disease specific) were included in the analysis. To further ascertain possible disease associations, we also calculated enrichment in DisGeNET (Table 4, Figure 5c), obtaining significance in schizophrenia (*1.38x10*$^{-6}$), autisic disorder (*6.67x10*$^{-7}$) and neurodevelopmental disorders (*5.72x10*$^{-10}$), to name a few. The most enriched tissue groups were the brain, blood vessel and nerve (Figure 5d).

The enriched pathways include statin inhibition of cholesterol production (*3.69x10*$^{-4}$); disruption of postsynaptic signalling by copy number variations (CNV) (*8.38x10*$^{-3}$); 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*$^{-3}$); and Reelin signalling (*8.99x10*$^{-3}$). A complete list of pathways enriched in the GSEA for maternal history, paternal history and parental meta-analysis can be found in Supplementary Table 5, and disease enrichment in Supplementary Table 6.

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). The binding of BDNF (neurotrophins family) to *NTRK2* (TRK family) initiates a cascade of intracellular signalling events including Rho protein signal transduction (GO:0035023) triggering neuron recognition (GO:0008038), amyloid-beta metabolic process (GO:0050435), cellular component of the Schaffer collateral - CA1 synapse (GO:0098685), Golgi to plasma membrane transport (GO:0006893) and ephrin receptor signalling pathway

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(GO:0048013). The results for GO enrichment of maternal and paternal histories can be found in Supplementary Figure 4 and Supplementary Table 7.

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[14], is not currently associated with any approved drug but received the highest tractability probability of 98.47%. It has a medical-quality pocket for small molecule development and given its importance in axon guidance (enriched pathway R-HSA-428542), could potentially be another therapeutic target to explore.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

3 Discussion

We applied GWANN to maternal and paternal histories of AD using data from the UKBB, followed by a parental meta-analysis. In doing so, we identified 23 known or nominated AD hits and 72 potentially novel hits. The association analysis was further supported by post-hoc enrichment analyses which highlighted enriched biological and disease pathways relevant to AD and neurodegeneration. Several GWANN hits were also identified to have known drugs that could possibly be repurposed, or a good probability for intervention with various modalities like small molecules and antibodies.

3.1 GWANN identifies hits associated with AD diagnosis

Of the genes that overlap between GWANN's discovery and discovery from previous GWAS, most of them have also been identified previously using familial history data in the UKBB. 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 UKBB. *GLIS3* was previously identified in a GWAS performed on cerebrospinal fluid tau levels[20]; *WWOX* was previously identified in the IGAP data[21]; and *IMMP2L* was previously identified for CNVs associated with AD[22]. Among the suggestive associations, other than the genes possibly linked to AD diagnosis, *EFNA5* was suggestively associated with psychosis in AD[15] and *MACROD2* with neurofibrillary tangles[16]. *ANK3*, a novel discovery in the latest GWAS conducted on the European Alzheimer's Disease BioBank (EADB)[19], was also identified by GWANN. Furthermore, observing the neighbourhood of the hits in the EADB analysis, we see a greater number of genes reaching nominal significance in the GWANN analysis as compared to the neighbourhood of genes that were insignificant in the EADB analysis (Supplementary Figure 5). The ability of GWANN to identify these genes from the UKBB data suggests the presence of subtle patterns in the data that were possibly missed by the traditional linear models but identified by our method.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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[23]. Two enriched pathways, R-HSA-428542 and R-HSA-8866376 are child-pathways of axon guidance. *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 its interaction with *ROBO1* (GWANN hit and AD target nomination) in the presence of *SLIT*, it 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[24]. 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[25].

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 the B cell depleting therapies have shown success in patients with disorders such as multiple sclerosis[26]. 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[27].

AD is often 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

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

synaptogenesis[28]. The disruption of the functioning of these genes due to CNVs affect synapse formation and neurodevelopment. They have also previously been associated with autism[29] and schizophrenia[30]. Another integral component of synaptic transmission is controlled by neurotrophins like *BDNF* (GWANN hit and nominated AD target), which play a very 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 major 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[31]. 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[32]. Pentoxifylline, a methylxanthine associated with *PDE1C*, has been suggested to improve cognitive function in patients with vascular dementia[33], reduce the odds of occurrence of Parkinson's disease[34] and reduce Aβ levels[34]. This further supports the quality of the novel GWANN hits and (i) suggests the possibility of

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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

Given that the method involves training a very 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[35]. 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 could be very 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 containing diagnosis of AD.

3.5 Conclusion

We applied our method to family history of AD using data from the UKBB, but it 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. The ability of GWANN to identify a set of possible targets that are part of meaningful biological pathways associated with AD and neurodegeneration, opens the world of GWAS to a new analysis method that could augment the success of existing methods in understanding the pathogenesis of AD and other diseases.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

4 Methods and Materials

4.1 Data source

We utilised data from the UKBB (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)[13,36]. We used imputed SNP data as input to GWANN. In addition, the covariates used were age (field 21003), sex (field 31), the first six genetic principal components (PCs) obtained from UKBB variables (field 22009), education qualification (field 6138) and a binary variable if this data was not available (1 if missing and 0 if not). Full UKBB cohort and variable descriptions are provided in the Supplementary material.

4.2 Cohort selection

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 were removed from the control groups[37]. 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 2 sexes (male and female) to obtain six broad groups — (age-group1, male), (age-group1, female) etc. In order 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.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 is 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 material.

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. 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

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 2500 bp upstream and downstream of it 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[38], a Python library built and optimised for machine learning and deep learning. Code will be available upon request.

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[39,40]. 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}.

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[41] to perform a gene-level meta-

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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.

4.7 Enrichment and post-hoc analyses

LDLink[42] was used to find the LD between SNPs using the "British in England and Scotland" population. 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, CERAD and COGDX). STRING V11.5[43] was used to perform PPI analysis of the genes that were significantly associated after the parental meta-analysis. Pathway enrichment was performed using GSEA[44]. 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[45]. We also performed an enrichment analysis of DEG sets in AD post-mortem brain regions — meta-analysis of AD brain transcriptomic data[17] and GSE118553[18] — using the same method, with the exception of multiple testing corrections. We applied a bonferroni correction for the number of DEG sets that were analysed. In addition to the GSEA, we performed a GO ORA of the GWANN hits using ClueGO v2.5.8[46] and a disease enrichment analysis with DisGeNET[47]. FUMA[48] was used to identify tissues enriched using the GWANN genes with P value<1x10-5. Finally, we used TargetDB[49] 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 material.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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; 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.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Author Contributions

U.G., W.S. and A.N.H. theorised and designed the method. W.S. carried out the pre-processing and quality control of the genomic data. U.G. developed the method and performed the experiments. U.G., W.S., L.W. and M.F. structured the post-hoc analyses. M.F. and U.G. carried out the different post-hoc analyses. M.F., D.N. and U.G. carried out the identification of drugs associated with the GWANN hits. B.U. performed the comparison of GWANN summary statistics with the EADB GWAS summary statistics. Q.L. helped with optimising the code pipeline for training the NNs. U.G., W.S., L.W. and M.F. wrote the manuscript. L.S., C.A., A.A., M.A., H.C and C.V.D. provided critical feedback that helped revise and finalise the manuscript. A.N.H. was the principal supervisor for the work carried out in this manuscript.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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. Goldstein, D. B. Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 1696–1698 (2009).
6. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
7. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
8. Marioni, R. E. *et al.* GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 99 (2018).
9. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
10. 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).
11. Rusk, N. Deep learning. *Nat. Methods* **13**, 35–35 (2016).
12. van Hilten, A. *et al.* GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. *Commun. Biol.* **4**, 1–9 (2021).

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

13. 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).
14. 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).
15. 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).
16. Dumitrescu, L. *et al.* Sex differences in the genetic predictors of Alzheimer's pathology. *Brain J. Neurol.* **142**, 2581–2589 (2019).
17. 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).
18. Patel, H. *et al.* Transcriptomic analysis of probable asymptomatic and symptomatic alzheimer brains. *Brain. Behav. Immun.* **80**, 644–656 (2019).
19. 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.
20. Cruchaga, C. *et al.* GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's disease. *Neuron* **78**, 256–268 (2013).
21. 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).
22. 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).
23. Zhang, L. *et al.* Roles and Mechanisms of Axon-Guidance Molecules in Alzheimer's Disease. *Mol. Neurobiol.* **58**, 3290–3307 (2021).
24. Rama, N. *et al.* Amyloid precursor protein regulates netrin-1-mediated commissural axon outgrowth. *J. Biol. Chem.* **287**, 30014–30023 (2012).

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

25. 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).
26. 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).
27. Kim, K. *et al.* Therapeutic B-cell depletion reverses progression of Alzheimer's disease. *Nat. Commun.* **12**, 2185 (2021).
28. Dean, C. & Dresbach, T. Neuroligins and neurexins: linking cell adhesion, synapse formation and cognitive function. *Trends Neurosci.* **29**, 21–29 (2006).
29. Glessner, J. T. *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459**, 569–573 (2009).
30. Walsh, T. *et al.* Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. *Science* **320**, 539–543 (2008).
31. 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).
32. 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).
33. 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).
34. Janitschke, D. *et al.* Methylxanthines and Neurodegenerative Diseases: An Update. *Nutrients* **13**, 803 (2021).
35. Hinton, G., Vinyals, O. & Dean, J. Distilling the Knowledge in a Neural Network. in *NIPS Deep Learning and Representation Learning Workshop* (2015).
36. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

37. Newby, D. *et al.* The relationship between isolated hypertension with brain volumes in UK Biobank. *Brain Behav.* e2525 doi:10.1002/brb3.2525.

38. 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).

39. Purcell, Shaun & Chang, Christopher. *PLINK 2.0*.

40. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).

41. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

42. 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).

43. 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).

44. 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).

45. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

46. 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).

47. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).

48. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

49. De Cesco, S., Davis, J. B. & Brennan, P. E. TargetDB: A target information aggregation tool and tractability predictor. *PloS One* **15**, e0232644 (2020).

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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. (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 UKBB 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

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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. (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); 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.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Tables

HGNC Symbol	Chromosome	P_{m+p}	P_m	P_p	HGNC Symbol	Chromosome	P_{m+p}	P_m	P_p
APOE	19	9.60E-190	1.00E-65	1.80E-158	PTPRN2	7	5.60E-09	1.90E-05	5.10E-05
TOMM40	19	9.60E-170	1.00E-65	2.40E-127	BCR	22	8.20E-09	1.70E-09	1.50E-01
APOC1	19	9.40E-153	1.00E-65	2.50E-102	MGAT4C	12	8.20E-09	1.10E-08	5.90E-02
BCAM	19	7.50E-75	1.00E-64	2.20E-14	PITPNM2	12	7.50E-08	6.10E-10	6.00E-01
GEMIN7	19	1.60E-21	2.60E-28	4.00E-01	CORO2B	15	1.30E-08	2.50E-07	1.10E-02
EXOC3L2	19	8.90E-08	3.50E-08	1.50E-01	KIF13A	6	1.60E-08	7.00E-09	1.20E-01
PPP1R37	19	1.90E-08	6.90E-08	3.70E-02	ASIC2	17	1.60E-08	1.80E-08	7.20E-02
GLIS3	9	4.10E-20	6.90E-08	8.40E-17	LMAN1	18	1.60E-08	1.10E-07	2.40E-02
ARSG	17	1.90E-15	1.30E-18	2.10E-01	BAHCC1	17	1.80E-08	2.80E-08	5.90E-02
C2CD2	21	7.70E-14	4.40E-09	3.30E-06	ELP4	11	1.80E-08	2.60E-05	1.30E-04
NFIA	1	5.40E-13	1.40E-14	1.20E-01	CNTN5	11	3.80E-07	4.90E-04	1.00E-04
PIGK	1	6.40E-08	6.90E-08	8.60E-02	DCC	18	2.10E-08	2.90E-06	2.00E-03
TENM2	5	1.80E-12	6.20E-06	5.00E-09	USP47	11	2.10E-08	4.40E-03	1.20E-08
EFNA5	5	2.60E-08	1.80E-08	9.70E-02	COL23A1	5	2.40E-08	7.10E-03	3.50E-09
SMAD9	13	1.90E-12	3.40E-16	5.40E-01	CPNE5	6	2.60E-08	1.80E-08	9.70E-02
SNRPB2	20	2.50E-12	4.80E-14	1.50E-01	SGCZ	8	2.80E-08	5.90E-07	1.10E-02
MTUS2	13	8.90E-12	1.70E-09	1.00E-03	PAX5	9	2.80E-08	1.30E-06	5.60E-03
MAGI2	7	8.90E-12	1.70E-09	1.00E-03	TANC1	2	2.80E-08	1.30E-06	5.60E-03
MYT1L	2	1.60E-11	1.10E-08	3.20E-04	LYPD6B	2	5.80E-07	2.90E-06	4.20E-02
PARD3B	2	1.90E-10	2.50E-07	1.90E-04	NRXN3	14	2.80E-08	2.60E-05	2.30E-04
CNTN4	3	2.70E-11	2.80E-08	2.30E-04	BDNF	11	3.00E-08	6.10E-06	1.40E-03
IQSEC1	3	4.70E-11	4.40E-09	2.00E-03	ARHGAP24	4	3.10E-08	4.40E-09	2.10E-01
ATP1B3	3	5.60E-09	2.70E-09	9.70E-02	EXT1	8	3.10E-08	2.00E-06	4.30E-03
ROBO1	3	2.10E-08	1.80E-08	8.60E-02	SORCS2	4	4.50E-08	7.40E-05	1.00E-04
FLNB	3	6.50E-08	9.00E-06	2.00E-03	SAMD4A	14	5.20E-08	1.80E-08	1.50E-01
RBMS3	3	9.30E-08	4.90E-04	1.40E-05	GPR137C	14	4.90E-07	6.90E-08	2.90E-01
ATP2C1	3	1.20E-07	5.90E-07	3.70E-02	TYW1B	7	5.90E-08	2.90E-06	5.60E-03
SERPINI1	3	1.70E-07	7.40E-05	5.50E-04	EXOC4	7	6.50E-08	8.80E-07	1.60E-02
EPHA6	3	1.90E-07	6.20E-06	8.60E-03	IMMP2L	7	1.60E-07	2.00E-04	1.30E-04
CADM2	3	2.30E-07	1.90E-05	3.60E-03	EPHA1-AS1	7	2.70E-07	2.20E-02	1.20E-08
CHCHD6	3	3.20E-07	2.00E-03	6.50E-06	ICAM3	19	6.40E-08	6.90E-08	8.60E-02
STAC	3	4.90E-07	2.60E-05	5.60E-03	DAB1	1	6.50E-08	9.00E-06	2.00E-03
HPSE2	10	3.00E-11	1.50E-11	5.20E-02	CARD11	7	7.70E-08	1.30E-05	1.70E-03
APH1B	15	3.80E-11	4.40E-09	1.70E-03	ARHGAP15	2	7.80E-08	2.60E-05	7.60E-04
MACROD2	20	4.80E-11	5.30E-05	1.20E-08	UBE2E2	3	5.40E-08	4.20E-06	3.60E-03
PDE1C	7	6.60E-11	6.10E-10	1.10E-02	UBE3A	15	1.30E-07	1.00E-09	6.30E-01
ARHGEF28	5	6.60E-11	6.90E-08	2.30E-04	C1QTNF9	13	1.30E-07	5.60E-08	1.50E-01
WWOX	16	1.30E-10	6.90E-08	4.50E-04	NTRK2	9	1.40E-07	2.90E-06	1.30E-02
TMEM170A	16	1.60E-07	6.90E-08	1.50E-01	BAIAP2L1	7	1.90E-07	1.10E-07	1.40E-01

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

HGNC Symbol	Chromosome	P_{m+p}	P_m	P_p	HGNC Symbol	Chromosome	P_{m+p}	P_m	P_p
LUZP2	11	1.90E-10	1.30E-06	2.70E-05	PDCL2	4	2.10E-07	2.00E-06	2.40E-02
FGGY	1	2.60E-10	2.20E-10	5.20E-02	NALCN	13	2.50E-07	7.10E-03	2.40E-07
ANK3	10	3.60E-10	2.80E-08	2.70E-03	RNF182	6	3.00E-07	4.90E-04	7.60E-05
CTNNA3	10	1.20E-07	5.30E-05	5.50E-04	CRISPLD1	8	4.10E-07	2.80E-08	3.70E-01
FSHR	2	6.80E-10	1.40E-02	3.50E-13	PEBP4	8	4.40E-07	3.20E-02	8.30E-09
NRXN1	2	3.40E-09	4.20E-06	1.90E-04	LRRC4C	11	4.90E-07	1.30E-05	1.10E-02
LRP1B	2	3.40E-08	1.90E-05	4.50E-04	WDR11-AS1	10	5.80E-07	5.90E-07	1.20E-01
VIPR2	7	2.80E-09	8.80E-04	1.90E-08	MYH11	16	6.30E-07	2.60E-05	7.30E-03
LINC00113	21	3.70E-09	7.70E-11	3.30E-01					

Table 1. GWANN hit genes. 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.

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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}

Table 2. Genes that had significant association with at least one of BRAAK, CERAD or COGDX in data obtained from the Agora AD knowledge portal (https://agora.adknowledgeportal.org/).

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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	**$8.77x10^{-7}$***	**$7.31x10^{-11}$***	**$5.82x10^{-4}$***	**$4.67x10^{-6}$***
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	**$1.49x10^{-3}$***	$2.24x10^{-2}$	$5.93x10^{-2}$

Table 3. (a) Enrichment P values of DEGs in four different brain regions for AD and other mental disorders analysed by Patel et al. [CITE] (b) Enrichment P values of DEGs in four different brain regions for the analysis between AD, asymptomatic AD and controls by Patel et al. [CITE]

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

medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Disease	FDR	GWANN hit count	Disease	FDR	GWANN hit count
Alzheimer's Disease	$2.42x10^{-05}$	38	Adolescent Idiopathic Scoliosis	$1.05x10^{-07}$	19
Schizophrenia	$1.38x10^{-06}$	37	Bipolar Disorder	$3.54x10^{-04}$	19
Mathematical Ability	$6.82x10^{-14}$	29	Systolic Pressure	$2.01x10^{-05}$	18
Obesity	$2.36x10^{-03}$	29	Autism Spectrum Disorders	$3.54x10^{-04}$	18
Body Mass Index	$2.55x10^{-11}$	28	Major Depressive Disorder	$4.49x10^{-03}$	17
Leukaemia, Myelocytic, Acute	$3.70x10^{-02}$	27	Impaired Cognition	$4.81x10^{-02}$	17
Intellectual Disability	$7.02x10^{-04}$	26	Abnormal Behaviour	$6.18x10^{-03}$	14
Autistic Disorder	$6.67x10^{-07}$	23	Familial (FPAH)	$2.26x10^{-02}$	14
Seizures	$8.28x10^{-03}$	23	Blood Protein Measurement	$3.78x10^{-02}$	14
Intelligence	$1.58x10^{-09}$	21	Attention Deficit Hyperactivity Disorder	$9.41x10^{-03}$	13
Neurodevelopmental Disorders	$5.72x10^{-10}$	20	Anxiety	$4.30x10^{-02}$	13
Global Developmental Delay	$1.56x10^{-02}$	20	Developmental Delay (Disorder)	$1.74x10^{-03}$	12
Body Height	$2.26x10^{-02}$	20	Alzheimer Disease, Late Onset	$3.28x10^{-04}$	11
Smoking	$2.94x10^{-11}$	19	Mental Disorders	$4.24x10^{-02}$	11
Scoliosis, Isolated, Susceptibility To, 3	$1.43x10^{-08}$	19	Dementia	$4.93x10^{-02}$	11

Table 4. Disease enrichment. Disease gene sets from DisGeNET that were significantly enriched after applying FDR correction. Column "GWANN hit count" indicates the number of genes that were part of the disease gene set among the 95 significant genes from GWANN.

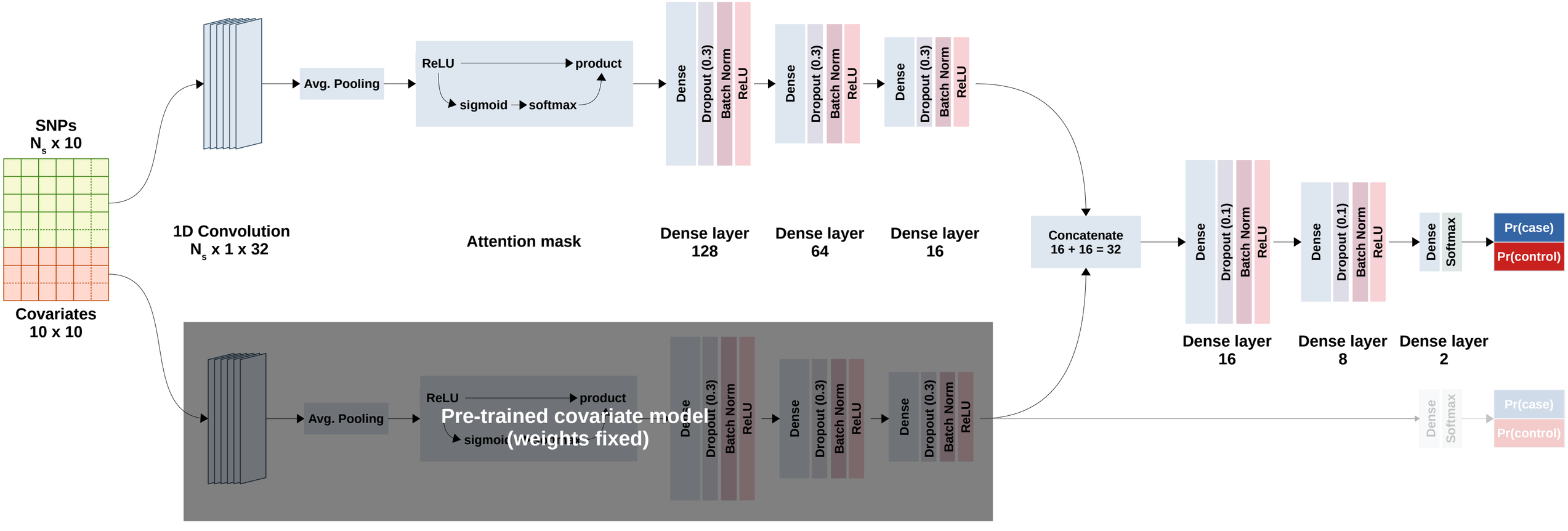
medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

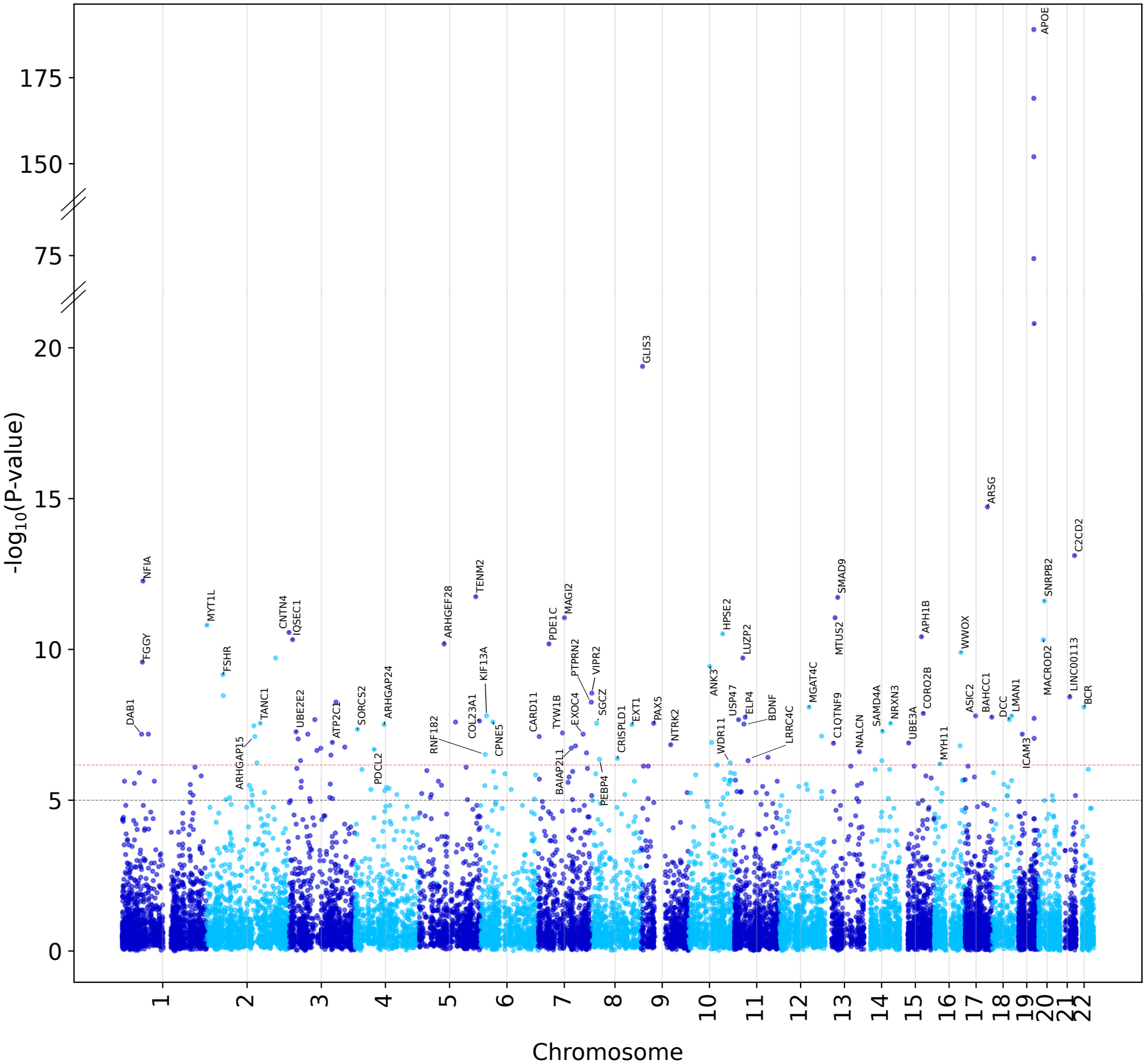
HGNC symbol	Drug Name	Type	Associated disease
BCR	Bosutinib	Small molecule	Cancer
	Dasatinib	Small molecule	Cancer
	Imatinib	Small molecule	Cancer
	Imatinib mesylate	Small molecule	Cancer
	Nilotinib	Small molecule	Cancer
	Nilotinib hydrochloride monohydrate	Small molecule	Cancer
	Ponatinib	Small molecule	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
	Selitrectinib	Small molecule	Keratitis, eye disease
PDE1C	Dipyridamole	Small molecule	Stroke
			Heart disease
			Bipolar disorder

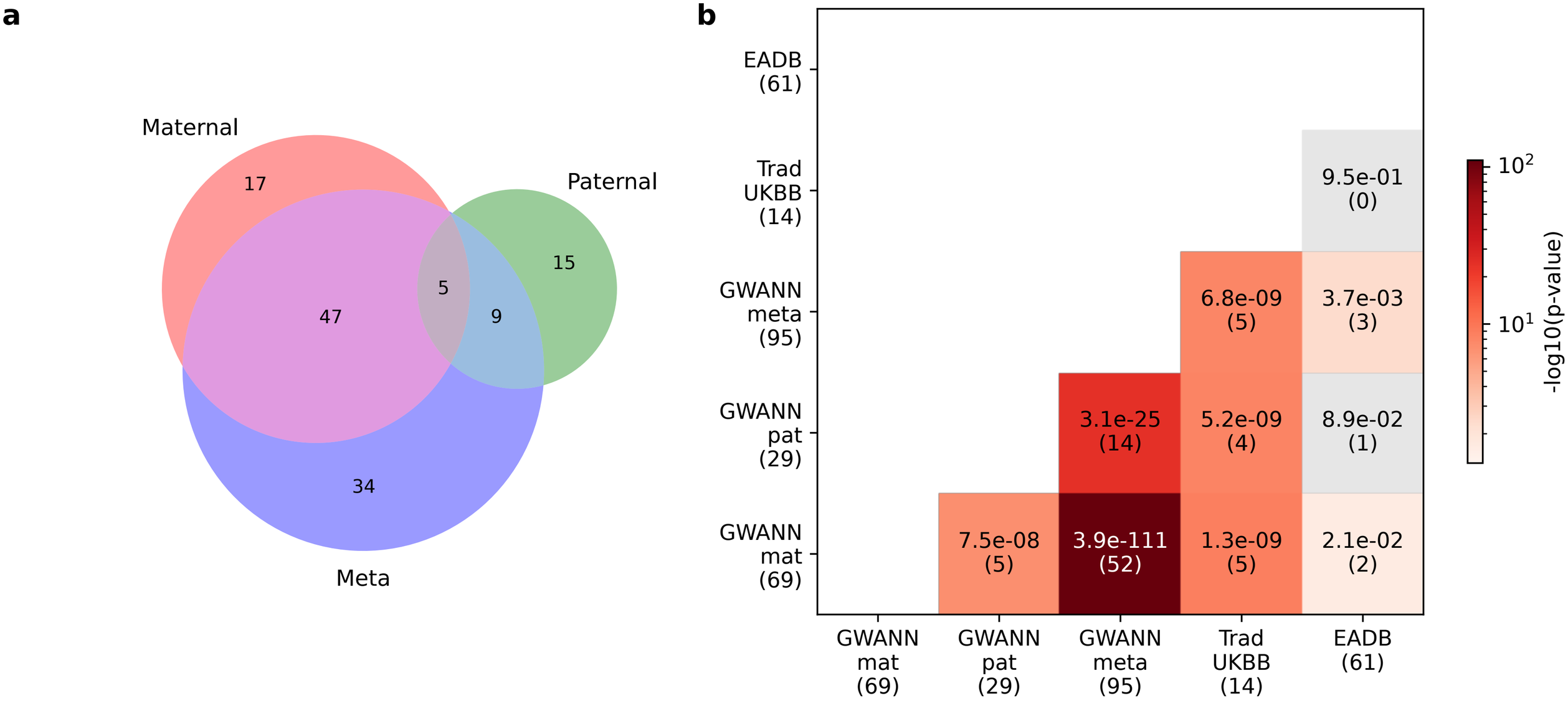
medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

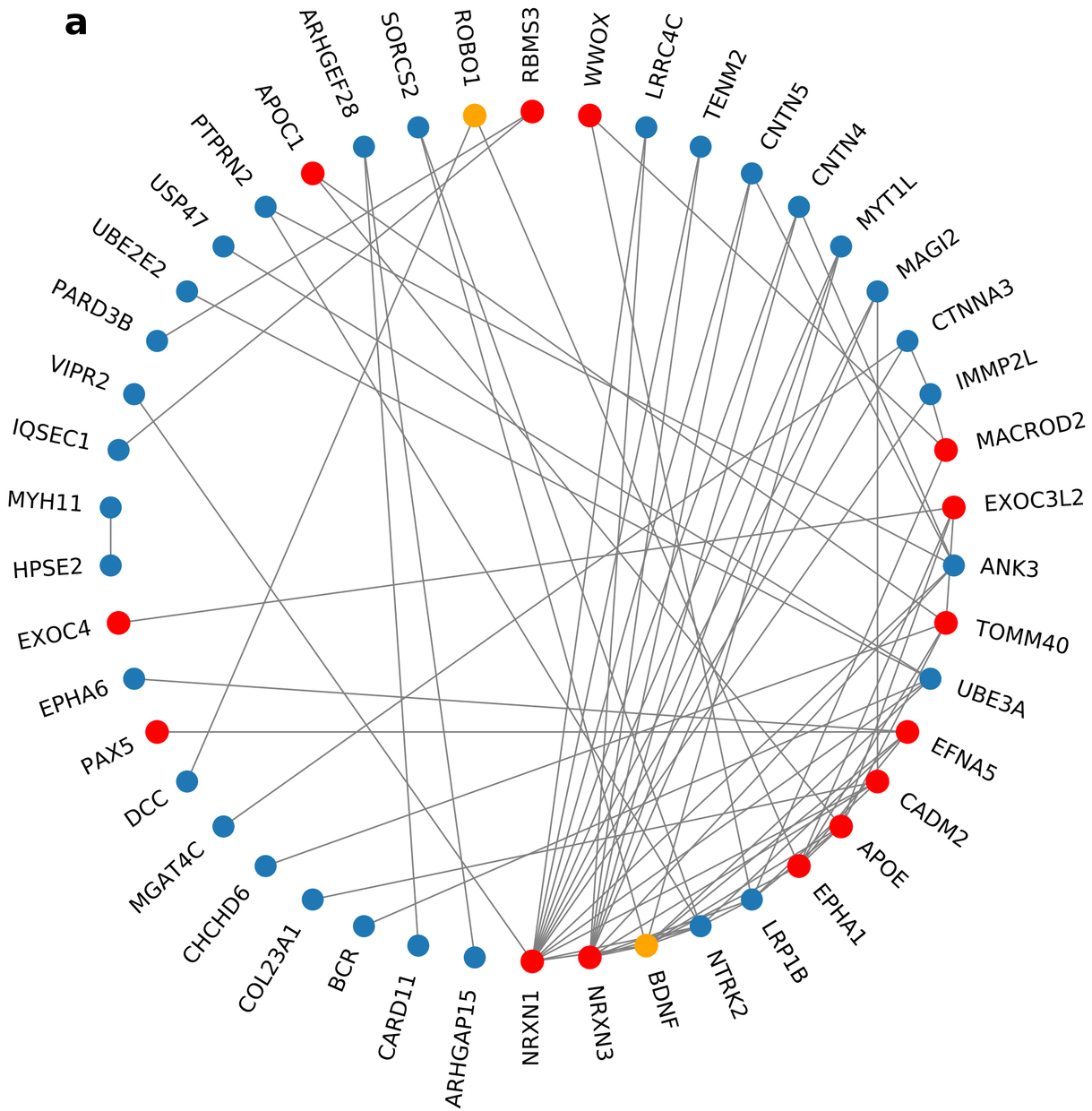
HGNC symbol	Drug Name	Type	Associated disease
			Diabetes mellitus
			COVID-19
	Pentoxifylline	Small molecule	Diabetes mellitus
			Brain disease
			Hepatic disease

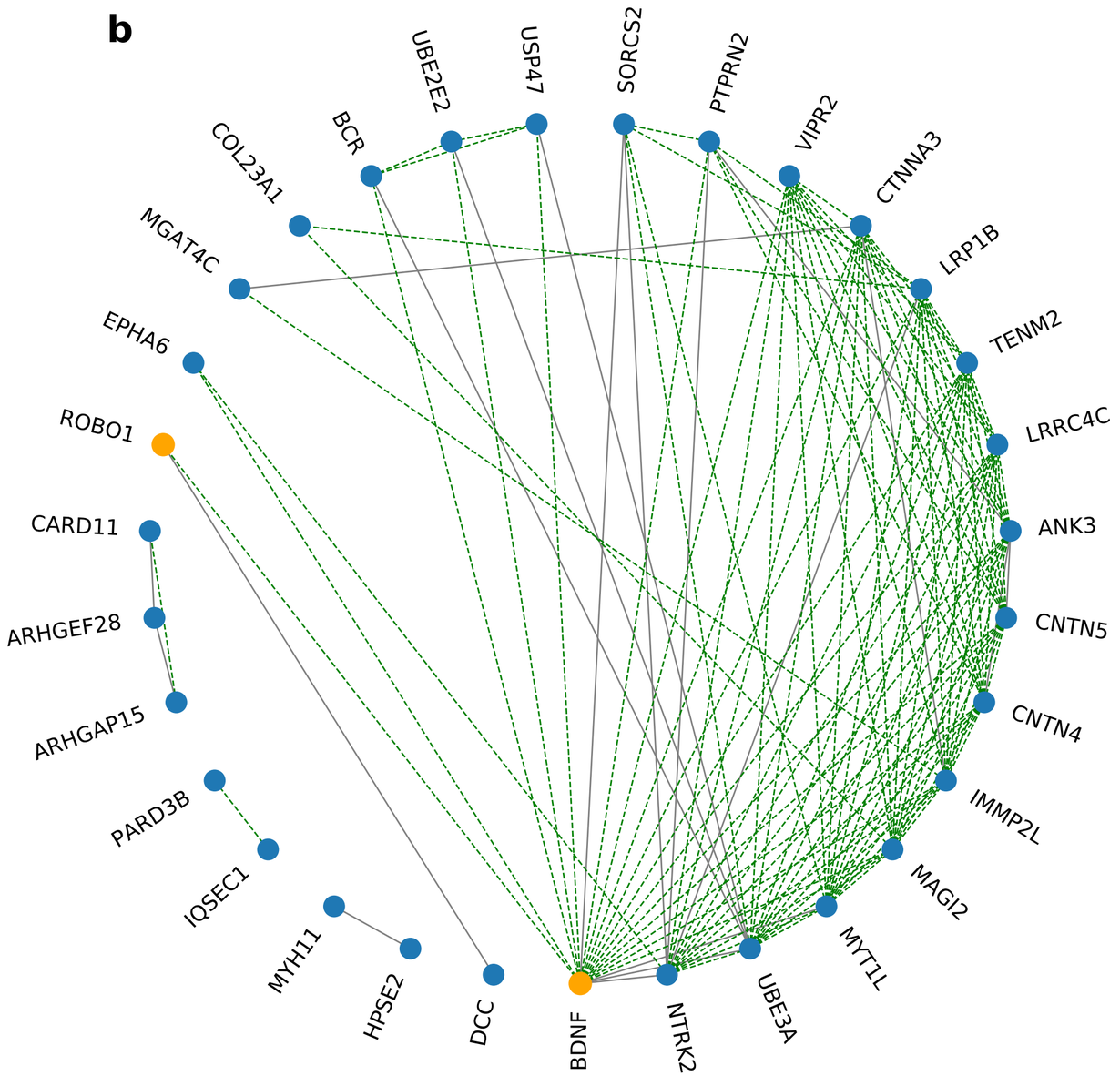
Table 5. Novel GWANN hits with known associated drugs and the main category of diseases they are used for.

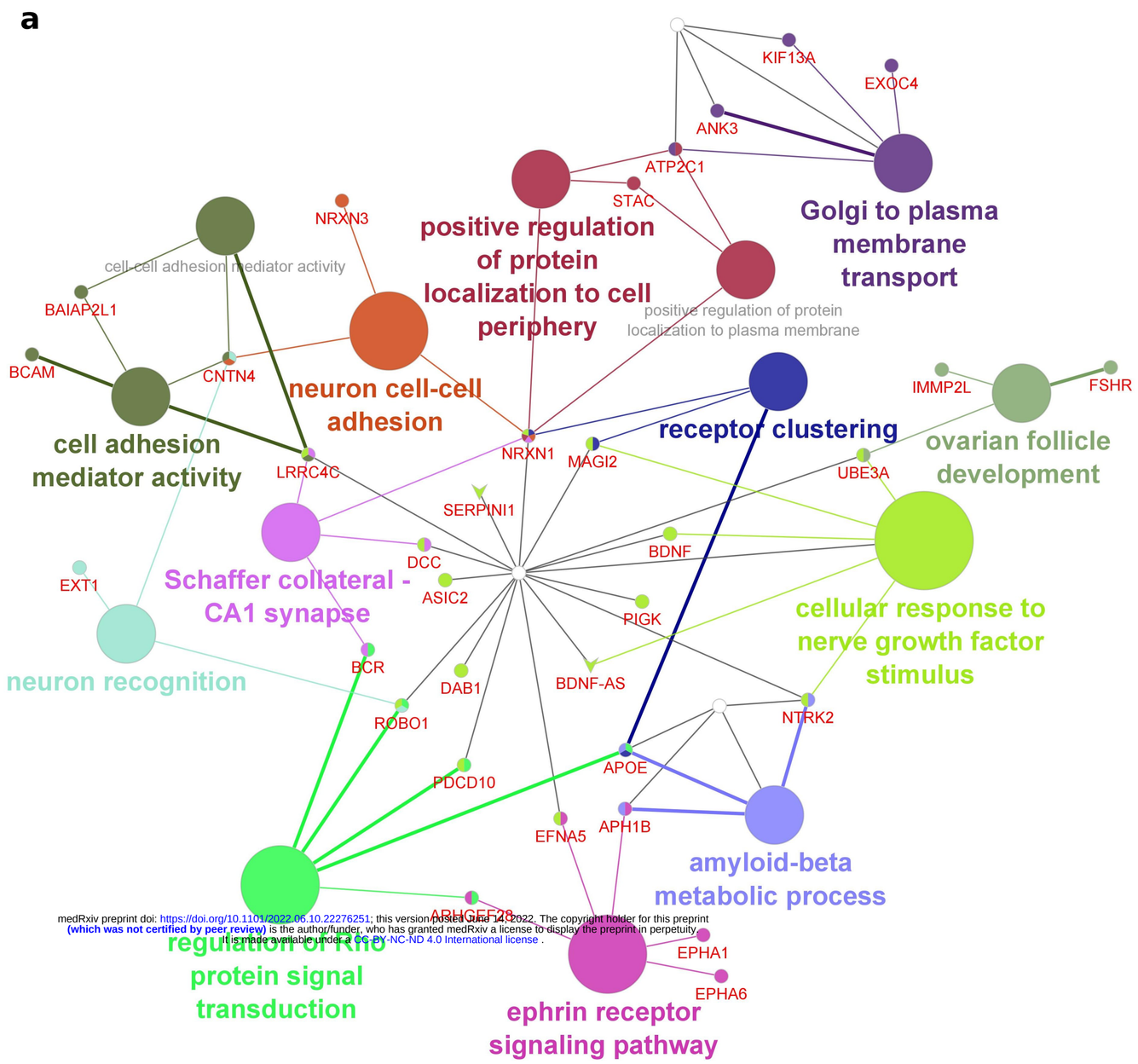
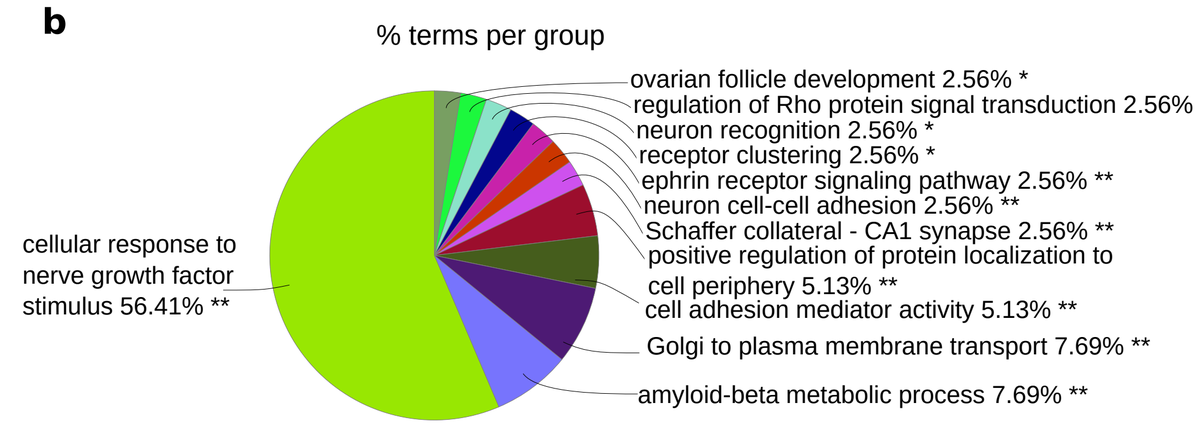
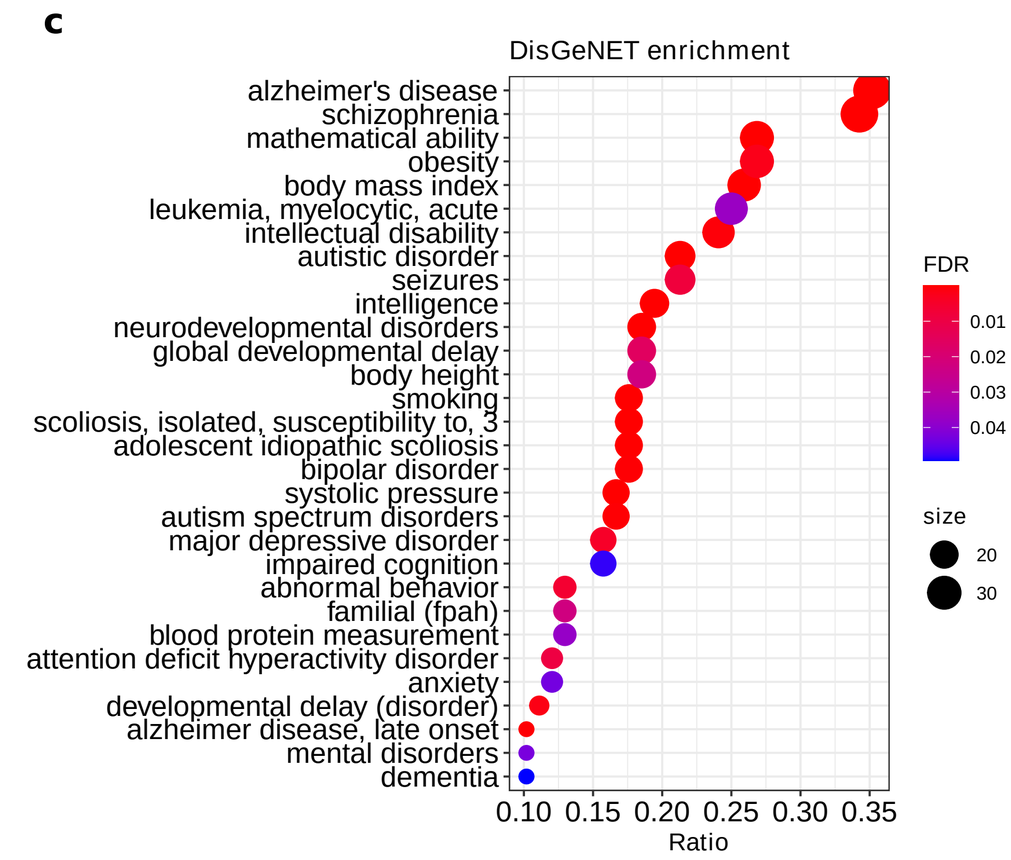
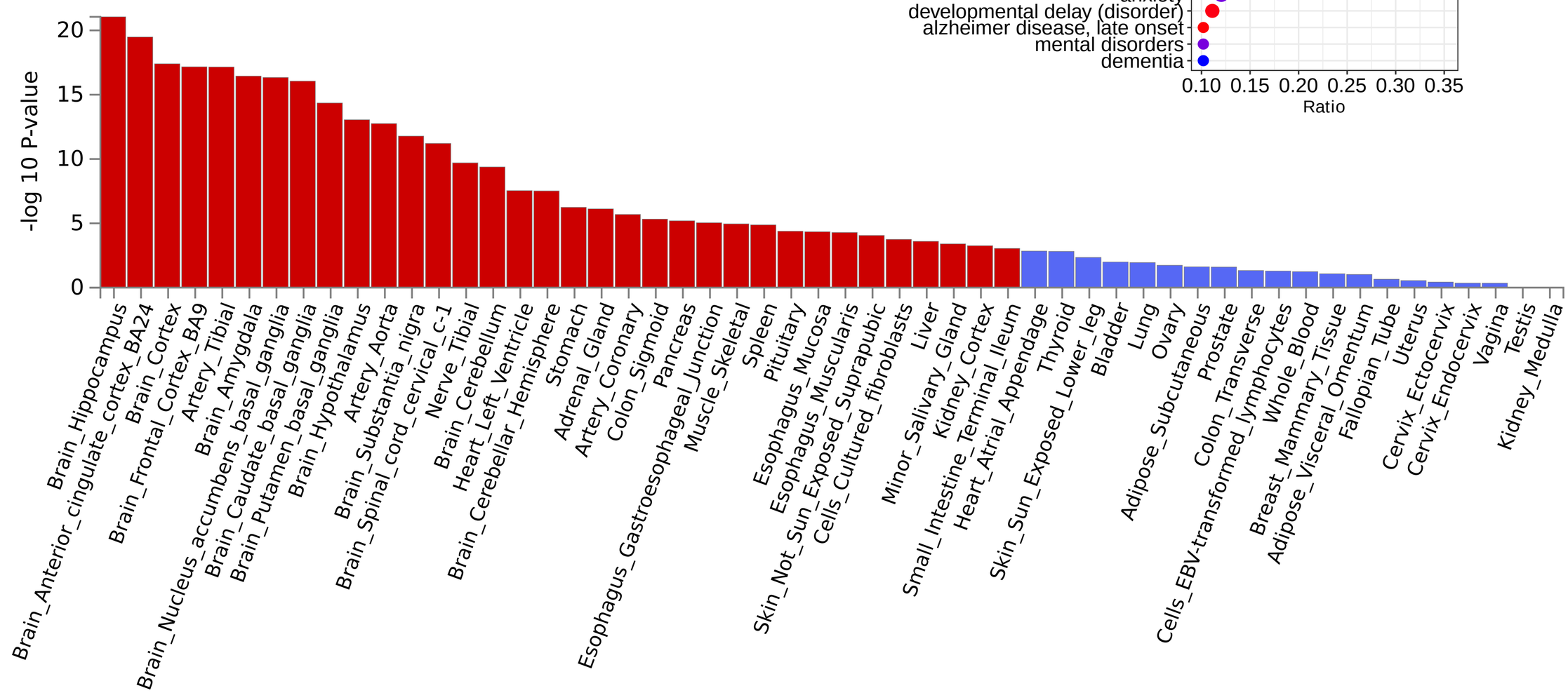

SNPs
N_s x 10
Covariates
10 x 10
1D Convolution
N_s x 1 x 32
Avg. Pooling
ReLU
product
sigmoid
softmax
Attention mask
Dense
Dropout (0.3)
Batch Norm
ReLU
Dense layer
128
Dense layer
64
Dense layer
16
Concatenate
16 + 16 = 32
Dense
Dropout (0.1)
Batch Norm
ReLU
Dense layer
16
Dense layer
8
Dense layer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


-log10(P-value)
175
150
75
20
15
10
5
0
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
APOE
GLIS3
ARSG
C2CD2
SNRPB2
MACROD2
LINC00113
BCR
NFIA
FGGY
DAB1
MYT1L
FSHR
ARHGAP15
TANC1
CNTN4
IQSEC1
UBE2E2
ATP2C1
SORCS2
PDCL2
ARHGAP24
RNF182
ARHGEF28
TENM2
COL23A1
KIF13A
CPNE5
CARD11
PDE1C
MAGI2
TYW1B
BAIAP2L1
EXOC4
PTPRN2
VIPR2
SGCZ
PEBP4
CRISPLD1
EXT1
PAX5
NTRK2
ANK3
WDR11
HPSE2
USP47
LUZP2
ELP4
BDNF
LRRC4C
MGAT4C
C1QTNF9
NALCN
SAMD4A
NRXN3
UBE3A
SMAD9
MTUS2
APH1B
CORO2B
MYH11
WWOX
ASIC2
BAHCC1
DCC
LMAN1
ICAM3


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
RBMS3
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
UBE2E2
USP47
SORCS2
PTPRN2
VIPR2
CTNNA3
LRP1B
TENM2
LRRC4C
ANK3
CNTN5
CNTN4
IMMP2L
MAGI2
MYT1L
UBE3A
NTRK2
BDNF
DCC
HPSE2
MYH11
IQSEC1
PARD3B
ARHGAP15
ARHGEF28
CARD11
ROBO1
EPHA6
MGAT4C
COL23A1
BCR



medRxiv preprint doi: https://doi.org/10.1101/2022.06.10.22276251; this version posted June 14, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .