

## **Whole-exome sequencing study identifies genes associated with Alzheimer's disease and related dementias**

Xiaoyi Raymond Gao<sup>1,2,3,4\*</sup>, Marion Chiariglione<sup>1</sup>, Alexander J. Arch<sup>1</sup>

<sup>1</sup>Department of Ophthalmology and Visual Sciences, The Ohio State University, Columbus, OH 43210, USA

<sup>2</sup>Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA

<sup>3</sup>Division of Human Genetics, The Ohio State University, Columbus, OH 43210, USA

<sup>4</sup>Ohio State University Physicians Inc., Columbus, OH, USA

\*Correspondence should be addressed to:

Xiaoyi Raymond Gao, PhD

Department of Ophthalmology and Visual Sciences

Department of Biomedical Informatics

Division of Human Genetics

The Ohio State University

Columbus, OH 43212

USA

Phone: (614) 293-5287

Fax: (614) 293-5602

Email: [raymond.gao@osumc.edu](mailto:raymond.gao@osumc.edu)

## Abstract

Alzheimer's disease (AD) is the most common form of late-onset neurodegenerative disease. Previous genome-wide association studies have identified numerous common genetic variants associated with AD. The contribution of rare variants to AD remains to be uncovered. AD-by-proxy, based on parental AD status, showed superior statistical power boost in recent AD studies. Using the UK Biobank (UKB) 368,865 whole-exome sequences of white British ancestry and AD-by-proxy (57,976 proxy AD cases and 310,889 non-AD proxy controls), we conducted the largest exome-wide association study of proxy AD to date. We identified 38 novel genes harboring rare variants for AD, such as *ANKRD36* ( $P = 6.02 \times 10^{-31}$ ), *MMP13* ( $P = 1.08 \times 10^{-6}$ ), *TUBA4A* ( $P = 7.15 \times 10^{-12}$ ), and *ZNF296* ( $P = 4.19 \times 10^{-14}$ ), demonstrating the power boost of aggregating rare variants and utilizing AD-by-proxy in gene discovery. We further replicated these genes in the FinnGen dataset. Notably, *MMP13* is a current drug target for AD and *TUBA4A* is a drug target for cognitive impairment in clinical trials. These results expand our knowledge of the genetic architecture of AD, especially the role of rare variants, and potential drug targets for AD.

## Introduction

Alzheimer's disease (AD) is the most common form late-onset neurodegenerative disease. It affects millions of people worldwide<sup>1</sup>. The age for late-onset AD is typically 65 and older. The heritability of AD is about 60-80%<sup>2</sup>. Previous genome-wide association studies (GWAS) and meta-analyses have uncovered common variants for AD with 75 genetic loci identified to date<sup>3</sup>. However, the role of rare variants in AD remains to be discovered. Well established genes harboring rare variants with association to AD only include *TREM2*, *SORL1*, and *ABCA7*<sup>4</sup>. Studying rare variants that are associated with AD will help advance our understanding of the biological mechanisms and potential new therapeutic targets for this devastating disease.

Large-scale biobanks have propelled recent rare-variant genetic association studies with unprecedented novel discoveries<sup>5-7</sup>. For example, the UK Biobank (UKB), a study of half a million individuals living in the United Kingdom with deep genotyping and genomic data<sup>8</sup>, represents one of the most successful biobanks in biomedical research. It recently released whole-exome sequences (WES) of 454,756 participants. A number of reports have used the UKB WES dataset for genetic association studies of quantitative traits and disease endpoints. However, for late-onset disease, e.g., AD, the number of cases can still be limited even in biobanks. To overcome this limitation, several recent studies used AD-by-proxy inferred from parental AD information and successfully made novel discoveries<sup>3,9</sup>.

The effectiveness of by-proxy phenotypes for GWAS was demonstrated by Liu and colleagues<sup>10</sup>. In a proxy case-control study design, proxy cases are the relatives of affected individuals<sup>10</sup>. Similarly, AD-by-proxy is to infer AD proxy case/control status based on parental AD history<sup>9</sup>. It is an approach that can significantly boost the effective sample size and statistical power of AD association tests<sup>9,10</sup>. The genetic correlation between AD-by-proxy and AD was found to be high (0.81)<sup>9</sup>. Hence, it is feasible to study AD through an AD-by-proxy approach. The robustness of this strategy in GWAS was also demonstrated

recently<sup>9</sup>. AD-by-proxy provides an effective approach to conduct AD association studies utilizing large-scale biobank resources to obtain substantial gain in statistical power<sup>9</sup>.

In this study, we carried out an exome-wide association study (ExWAS) to identify genes associated with proxy AD using the UKB WES data. We further conducted replication of the identified genes in AD phenotypes in the FinnGen dataset, a biobank of over 200,000 Finnish people. To the best of our knowledge, this study represents the largest ExWAS of AD-by-proxy so far. Our results uncovered novel genes harboring rare variant with association to AD and advanced our understanding of the biological mechanisms of the disease. We further explored potential drug targets of the identified genes.

## Results

In this AD analysis using the AD-by-proxy phenotype, a total of 368,865 participants of white British ancestry (57,976 cases and 310,889 controls) from UKB were included following Jansen et al.'s proxy AD definition<sup>9</sup>. The mean (standard deviation) age was 59.2 (6.9) years for the cases and 56.7 (8.1) years for the controls. Among the AD cases, 33,128 (57.1%) were female, and among the non-AD controls, 172,152 (55.4%) were female.

In this ExWAS analysis, we replicated 26 previously known AD genes, including all established genes harboring rare variants (*TREM2*, *SORL1*, and *ABCA7*) and identified 38 novel potential AD-by-proxy genes ( $P < 2.5 \times 10^{-6}$ ) from our gene-based analysis using SAIGE<sup>11</sup>, a state-of-the-art tool for gene-based association testing. The genomic control lambda was 1.0, indicating our analysis is properly controlled for population structure and cryptic relatedness. The corresponding quantile-quantile plot is shown in Supplementary Figure 1. Table 1 shows the list of gene-based associations with AD-by-proxy. Genes were considered novel if they were not found associated with AD within the GWAS catalog or the Alzheimer's Disease Variant Portal (ADVP) databases (see web resources). *ANKRD36* shows a highly significant association with proxy AD ( $P = 6.02 \times 10^{-31}$ ), followed by other hits, such as *PRB4* ( $P = 7.78 \times$

$10^{-15}$ ), *ZNF296* ( $P = 4.19 \times 10^{-14}$ ), *TTK* ( $2.10 \times 10^{-12}$ ), *TUBA4A* ( $P = 7.15 \times 10^{-12}$ ), *TUBA4B* ( $P = 7.98 \times 10^{-12}$ ), and *MMP13* ( $P = 1.08 \times 10^{-6}$ ). We further queried the FinnGen website for the association of the identified genes with AD. Many of these genes were identified to have strong associations with AD in FinnGen. For example, there is a highly significant association with AD for *ZNF296* ( $P = 8.9 \times 10^{-100}$ ), *ADAM8* ( $P = 2.3 \times 10^{-6}$ ), *HIGD1B* ( $P = 6.6 \times 10^{-6}$ ), *EFTUD2* ( $P = 6.6 \times 10^{-6}$ ) and *TTK* ( $P = 1.4 \times 10^{-5}$ ). A Manhattan plot of the gene-based p-values is shown in Figure 1.

To seek biological support for the identified genes, we used the gene expression and differential expression information in the Genevestigator<sup>12</sup> and the Agora databases, respectively. Supplementary Figure 2 displays the bulk tissue gene expression information from Genevestigator. Most of the identified novel genes, such as *ANKRD36*, *TUBA4A*, *TUBA4B*, *MMP13*, *MUC16*, *TTK*, and *ZNF296*, are expressed in the brain. Figure 2 shows the differential expression (DE) evidence for these identified genes from the Agora knowledge portal, a platform for exploring AD genes, including DE, initially developed by the National Institute on Aging's Accelerating Medicines Partnership in Alzheimer's Disease (AMP-AD) consortium. In the figure, the diameter and the color of the circles indicate p-value and log<sub>2</sub> fold change, respectively. Many of the significant DE genes are perturbed in parahippocampal gyrus and temporal cortex. Several of identified genes, such as *RASGRF1*, *HIGD1B*, and *LLGL2*, show significant DE across most of the AD-related brain regions.

The novel genes also included some promising drug targets for AD. To query potential drug targets, we used the online Open Targets tool. Several of the identified genes associated with proxy AD have known drug targets in clinical trials, as shown in Table 2. The genes *TUBA4A* is targeted by Davunetide, a tubulin stabilizer, which may serve as a possible treatment for cognitive impairment. The drug Doxycycline, a matrix metalloproteinase inhibitor, targets the gene *MMP13* and is being researched for use against AD. *MUC16* and *TTK* are targeted by Oregovomab and BAY-1217389 for treating ovarian cancer and cancer, respectively.

## Discussion

In this study, we conducted the largest ExWAS of proxy AD to date using the WES data from UKB. By utilizing an AD-by-proxy approach, we have expanded our knowledge of the role of rare variants for AD and related dementias. We identified 38 novel genes for proxy AD, demonstrating the power of AD-by-proxy and rare-variant aggregation in gene discovery. We also replicated 26 previously reported AD genes, including all the well-established genes harboring rare variants, i.e., *TREM2*, *SORL1*, and *ABCA7*, further demonstrating the robustness of our approach. Of note, among the novel genes, *MMP13* and *TUBA4A*, have already been targeted in clinical trials for AD and cognitive impairment.

AD gene mapping can be hindered by the limited number of cases available, especially for rare-variant investigation. Previous approaches have used meta-analysis, which has been successful in identifying numerous common variants associated with AD. Another approach is to use AD-by-proxy<sup>9</sup> taking advantage of the enormous resources available in large-scale biobanks, e.g., UKB. This disease-by-proxy approach has been used in several recent studies on GWAS datasets<sup>3,9,13</sup>. Here, we demonstrated its effectiveness using the UKB WES data and identified novel genes for AD.

The identified ExWAS significant genes show many interesting patterns. The most significant one, *ANKRD36* ( $P = 6.01 \times 10^{-31}$ ), is highly significantly associated with autoimmune disease in FinnGen (Beta 8,  $P = 9.6 \times 10^{-10}$ ). A recent report showed that AD may be rooted in an autoimmune response<sup>14</sup>. *ANKRD36* has also been reported for an association with schizophrenia<sup>15</sup>, a severe psychiatric illness that may have a white matter deficit connection with AD<sup>16</sup>. *ANKRD36* is also involved in hypertension<sup>17</sup>, a risk factor for AD. Many genes, such as *CEACAM19*, *BCL3*, *BCAM*, *NECTIN2*, *TOMM40*, *APOE*, *CLASRP*, and *OPA3*, are in the about 1M region around *APOE* on chromosome 19 and show both significant common- and rare-variant associations with AD. We further identified two other genes, i.e., *MIR8085* ( $P = 1.09 \times 10^{-15}$ ) and *ZNF296* ( $P = 4.19 \times 10^{-14}$ ), that showed significant rare-variant

associations with proxy AD in the 1M region around *APOE*. Many of the other significant genes, such as *DAB1*, *OMAI*, *PVR*, *CEACAM19*, *BCL3*, *CLASRP*, and *ZNF296*, are highly significantly associated with statin medication in FinnGen (online Beta 8 results), which may indicate the interrelationship between AD and cardiovascular health.

There is a dearth of drugs available for AD. With the goal of conquering AD by 2025<sup>18</sup>, we clearly need more candidate genes for AD drug targets. Genes in our list has already given promising results. For example, *MMP13* ( $P = 1.07 \times 10^{-6}$ ) in our novel gene list is an AD drug target for Doxycycline and has a total of 4 drugs with 103 indications. *TUBA4A* is in clinical trials for cognitive impairment and has a total of 19 drugs with 69 indications. *MUC16* and *TTK* are known targets for some cancer drugs, which may possibly be repurposed for AD. The other genes in our list may serve as novel drug targets as well.

Our study is not without limitations. AD-by-proxy is a proxy for AD based on parental AD history and is less defined than clinically diagnosed AD. However, due the high heritability of AD, the case status for offspring of AD parents is expected to be high. Coupled with large sample size of biobanks, like UKB, proxy AD can substantially increase the statistical power of AD discoveries. Moreover, AD-by-proxy has been proven successful in AD discoveries in previous AD GWAS studies<sup>3,9</sup>. Using this approach, we not only identified novel AD genes but also confirmed 26 known ones. Most of the novel ones showed DE for AD in the Agora resources. Several of these novel genes identified are also drug targets for AD in clinical trials, which further demonstrated the effectiveness of our AD-by-proxy approach in AD genetic discoveries. Nevertheless, replication of our results in clinical diagnosed AD in other datasets and diverse populations is needed. There are many different ways to analyze rare variants, e.g., including synonymous variants<sup>19</sup>, sliding window<sup>20</sup>, and all inclusive<sup>21</sup>. To our knowledge, there is no single rare-variant analysis method that performs the best in all situations. Using other approaches may identify more novel AD genes.

In conclusion, we carried out the largest ExWAS of AD-by-proxy to date. In addition to improving our understanding of the genetic architecture of AD and related dementias, especially the role of rare variants, the results of our study may serve as a valuable resource for exploring potential drug targets for AD given the relevance of *MMP13* and *TUBA4A*, all of which are in our novel gene list, as drug targets for AD and cognitive impairment.

## **Materials and Methods**

### **UKB Resource**

UKB is a large prospective cohort study comprising half a million participants living in the United Kingdom. The details of the UKB cohort have been previously reported. Briefly, participants aged 40 to 70 years and registered with the National Health Services were enrolled during the baseline recruitment from 2006 to 2010. Family history, lifestyle information, self-reported health information, electronic medical records, and DNA samples were collected. Approximately 94% of the participants reported their ethnic background as white, and we used individuals of white British ancestry for this study. Our access to the UKB data was obtained under application number 23424, and we obtained fully de-identified data.

### **FinnGen Resource**

FinnGen is a large Finnish biobank that focuses on the Finnish population. It has enrolled over 200,000 participants, who are aged 18 or older and have a median age of 63<sup>22</sup>. By the end of 2023, it is projected to have more than half a million participants<sup>22</sup>. Comprehensive genotype and phenotype data have been generated and collected since 2017. Genotyping was done using the custom Axiom FinnGen1 array as well as legacy arrays. The genotypes were further imputed to 17 million markers based on whole-genome sequences of Finns. Phenotypes were primarily constructed using International Classification of Disease Ninth and Tenth Revision (ICD-9 and ICD-10) codes. FinnGen GWAS summary statistics for various phenotypes are publicly available (see web resources).



## UKB WES and Quality Control

The WES data for UKB participants were generated at the Regeneron Genetic Center<sup>5,23</sup>. Previously, details on the sequencing, variant calling, and quality control processes have been reported<sup>6,23</sup>. In brief, the Illumina NovaSeq 6000 platform was used for sequencing with 75 base pair paired-end reads, and the SPB protocol was utilized for variant calling and quality control<sup>24</sup>. The WES data was of high quality, with over 20x coverage of 95.8% of targeted bases. We overlaid the data with our derived proxy AD phenotype and retained samples with missing rates less than 2.5% and autosomal variants with minor allele frequencies (MAFs) less than 0.01, call rates greater than 95%, and minor allele counts greater than or equal to 1 (19.6 million rare variants). We used VEP<sup>25</sup> and annovar<sup>26</sup> for annotating the variants.

## AD-by-proxy phenotype in UKB

We derived the AD-by-proxy phenotype following Jansen *et al.*'s previous report on proxy AD<sup>9</sup> based on the AD information obtained through the intake UKB questionnaire and electronic health records. “By proxy” indicates that the phenotype built includes information about the status of the immediate biological family of the individual regarding the illness studied. The phenotype also includes the status of the individual itself regarding whether they have AD or not. During the intake process with UKB, each individual was asked about their biological mother's and father's illnesses and whether they suffer from AD (data fields 20107 and 20110 for the illnesses of father and mother, respectively) as well as the current parent's age (data fields 2946 and 1845) or age when they passed away if applicable (data fields 1807 and 3526). Based on these fields, if an individual's parent was found to have AD he/she was given a “risk” score of 1, on the other hand, if they were found to be unaffected, their score was computed using their age or age at death to consider the fact that they might not have been old enough for a diagnosis of this late-onset disease. The formula used for this is the proportion of the parent's age to age 100:  $parental\_risk\_score = (100 - age) / 100$ . This parental risk score for the unaffected parent, was also capped at 0.32, corresponding to a risk similar to that of the general occurrence of AD in the population<sup>27</sup>. We used the International Classification of Disease Tenth Revision (ICD-10) codes (data

fields 41202 and 41204 for diagnoses - main and secondary, respectively) to characterize a participant's own illnesses information on AD. Participants with the codes G30 (AD) or F00 (Dementia in AD) were considered cases and were given the maximum "risk" score possible of 2 regardless of the AD status of their parents. The resulting AD-by-proxy score (including biological parents' information) is between 0 and 2 with 0 representing both parents being unaffected and 2 when both parents or the individual were affected. Anyone with a score over 1 was considered a case. We also performed quality control on this phenotype and removed individuals who answered "do not know" or "prefer not to answer" on the questionnaires about their parents' illnesses, individuals without any parents' age, outliers for heterozygosity, individuals with ten or more third-degree relatives, and individuals with sex chromosome aneuploidy. For this study, we restricted our analysis to self-identified white British (UKB data field 21000) participants and removed outliers with genetic ancestry at least six standard deviations from the means of the first two principal components. This resulted in 57,976 AD cases and 310,889 non-AD controls (white participants) after overlapping with the WES data.

### **Gene-based ExWAS analysis**

We used SAIGE<sup>11</sup>, a generalized mixed model approach that can adjust for both genetic relationship and population stratification, to carry out our gene-based association tests. SAIGE includes several rare variant collapsing and aggregation tests, such as SKAT-O<sup>28</sup>, burden<sup>29</sup>, and SKAT<sup>30,31</sup>. We used the SKAT-O approach, which adaptively combines the strengths of both the burden and SKAT test statistics and demonstrates advantages in statistical power and robustness<sup>32</sup>. We included predicted loss of function (pLOF) variants, such as stop gained, stop lost, start lost, splice donor, splice acceptor, and frameshift, based on the VEP<sup>25</sup> annotation and gnomAD pLOF variants<sup>33</sup>, as well as missense variants (MAF < 0.005). We adjusted for age, sex, and the first 10 principal components of genetic ancestry. Genes with  $P < 2.5 \times 10^{-6}$  were declared ExWAS significant.

### **AD lookup in FinnGen**

To replicate our significant genes, we queried the FinnGen<sup>22</sup> online GWAS summary statistics (Beta 8, <https://www.finnngen.fi/>) and checked their associations with AD phenotypes in the FinnGen database. There are 14 endpoint AD phenotypes in FinnGen. We queried the Alzheimer's disease (wide definition), which is defined as F00\*, G30 (ICD10 hospital discharge or cause of death), 3310 (ICD9 hospital discharge or cause of death), 29010 (ICD8 hospital discharge or cause of death), 307 (KELA codes), and N06D (ATC codes), and dementia in Alzheimer disease, which is defined as F00\* (ICD10 hospital discharge or cause of death) and 29010 (ICD8 hospital discharge or cause of death) based on the FinnGen health registries (<https://risteys.finnngen.fi/>).

### **Gene Expression Evidences**

To query bulk RNA gene expression information in brain tissues, we used Genevestigator<sup>12</sup>, an online gene expression database. Expression profiles were obtained from 2,041 human nervous system tissues and displayed as box plots indicating low, medium, or high expression for the corresponding specific tissue. To query differentially expressed genes, we used Agora (see web resources), a web-based application that hosts evidence for whether or not genes are associated with AD.

### **Drug Targets Prioritization**

To identify potential drug targets for the genes identified in our analyses, we used the Open Targets<sup>34</sup> online resource. We queried the website for known drugs, mechanisms of action (source ChEMBL), and disease information related to AD. This drug information provides valuable insights into the relevance of the identified genes to AD and potential drugs for repurposing.

## **Acknowledgements**

This work was supported in part by grants from the National Institute on Aging, National Institutes of Health (NIH; Bethesda, MD, USA): RF1AG060472. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We would like to thank the study participants and investigators from the UK Biobank as well as the staff who aided in data collection and processing. We also want to acknowledge the participants and investigators of the FinnGen study.

## **Conflict of Interest Statement**

None declared.

### *Web Resources*

The URLs for downloaded data and programs:

Agora, <https://agora.adknowledgeportal.org>

Alzheimer's Disease Variant Portal (ADVP), <https://advp.niagads.org>

ANNOVAR, <http://annovar.openbioinformatics.org/>

ChEMBL, <https://www.ebi.ac.uk/chembl/>

GWAS Catalog, <https://www.ebi.ac.uk/gwas/>

FinnGen, <https://www.finngen.fi/>

SAIGE, <https://github.com/weizhouUMICH/SAIGE>

UK Biobank, <https://www.ukbiobank.ac.uk>

VEP, <https://useast.ensembl.org/info/docs/tools/vep/index.html>

R, <https://www.r-project.org>

## References

1. 2022 Alzheimer's disease facts and figures. (2022). *Alzheimers Dement* 18, 700-789. 10.1002/alz.12638.
2. Gatz, M., Pedersen, N.L., Berg, S., Johansson, B., Johansson, K., Mortimer, J.A., Posner, S.F., Viitanen, M., Winblad, B., and Ahlbom, A. (1997). Heritability for Alzheimer's Disease: The Study of Dementia in Swedish Twins. *The Journals of Gerontology: Series A* 52A, M117-M125. 10.1093/gerona/52A.2.M117.
3. Bellenguez, C., Kucukali, F., Jansen, I.E., Kleinedam, L., Moreno-Grau, S., Amin, N., Naj, A.C., Campos-Martin, R., Grenier-Boley, B., Andrade, V., et al. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet* 54, 412-436. 10.1038/s41588-022-01024-z.
4. Khani, M., Gibbons, E., Bras, J., and Guerreiro, R. (2022). Challenge accepted: uncovering the role of rare genetic variants in Alzheimer's disease. *Mol Neurodegener* 17, 3. 10.1186/s13024-021-00505-9.
5. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C., Liu, D., Locke, A.E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* 599, 628-634. 10.1038/s41586-021-04103-z.
6. Wang, Q., Dhindsa, R.S., Carss, K., Harper, A.R., Nag, A., Tachmazidou, I., Vitsios, D., Deevi, S.V.V., Mackay, A., Muthas, D., et al. (2021). Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* 597, 527-532. 10.1038/s41586-021-03855-y.
7. Nag, A., Middleton, L., Dhindsa, R.S., Vitsios, D., Wigmore, E., Allman, E.L., Reznichenko, A., Carss, K., Smith, K.R., Wang, Q., et al. (2021). Assessing the contribution of rare-to-common protein-coding variants to circulating metabolic biomarker levels via 412,394 UK Biobank exome sequences. medRxiv, 2021.2012.2024.21268381. 10.1101/2021.12.24.21268381.
8. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203-209. 10.1038/s41586-018-0579-z.
9. Jansen, I.E., Savage, J.E., Watanabe, K., Bryois, J., Williams, D.M., Steinberg, S., Sealock, J., Karlsson, I.K., Hagg, S., Athanasiu, L., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet* 51, 404-413. 10.1038/s41588-018-0311-9.
10. Liu, J.Z., Erlich, Y., and Pickrell, J.K. (2017). Case-control association mapping by proxy using family history of disease. *Nat Genet* 49, 325-331. 10.1038/ng.3766.
11. Zhou, W., Zhao, Z., Nielsen, J.B., Fritsche, L.G., LeFaive, J., Gagliano Taliun, S.A., Bi, W., Gabrielsen, M.E., Daly, M.J., Neale, B.M., et al. (2020). Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat Genet* 52, 634-639. 10.1038/s41588-020-0621-6.
12. Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W., and Zimmermann, P. (2008). Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Adv Bioinformatics* 2008, 420747. 10.1155/2008/420747.
13. Marioni, R.E., Harris, S.E., Zhang, Q., McRae, A.F., Hagenaars, S.P., Hill, W.D., Davies, G., Ritchie, C.W., Gale, C.R., Starr, J.M., et al. (2018). GWAS on family history of Alzheimer's disease. *Transl Psychiatry* 8, 99. 10.1038/s41398-018-0150-6.
14. Meier-Stephenson, F.S., Meier-Stephenson, V.C., Carter, M.D., Meek, A.R., Wang, Y., Pan, L., Chen, Q., Jacobo, S., Wu, F., Lu, E., et al. (2022). Alzheimer's disease as an autoimmune disorder of innate immunity endogenously modulated by tryptophan metabolites. *Alzheimers Dement (N Y)* 8, e12283. 10.1002/trc2.12283.
15. Goes, F.S., McGrath, J., Avramopoulos, D., Wolyniec, P., Pirooznia, M., Ruczinski, I., Nestadt, G., Kenny, E.E., Vacic, V., Peters, I., et al. (2015). Genome-wide association study of

- schizophrenia in Ashkenazi Jews. *Am J Med Genet B Neuropsychiatr Genet* 168, 649-659. 10.1002/ajmg.b.32349.
16. Kochunov, P., Zavaliangos-Petropulu, A., Jahanshad, N., Thompson, P.M., Ryan, M.C., Chiappelli, J., Chen, S., Du, X., Hatch, K., Adhikari, B., et al. (2021). A White Matter Connection of Schizophrenia and Alzheimer's Disease. *Schizophr Bull* 47, 197-206. 10.1093/schbul/sbaa078.
  17. Yan, Y., Wang, J., Yu, L., Cui, B., Wang, H., Xiao, X., Zhang, Y., Zheng, J., Wang, J., Hui, R., and Wang, Y. (2021). ANKRD36 Is Involved in Hypertension by Altering Expression of ENaC Genes. *Circulation research* 129, 1067-1081. 10.1161/CIRCRESAHA.121.319883.
  18. Vradsenburg, G. (2015). A pivotal moment in Alzheimer's disease and dementia: how global unity of purpose and action can beat the disease by 2025. *Expert Rev Neurother* 15, 73-82. 10.1586/14737175.2015.995638.
  19. Karczewski, K.J., Solomonson, M., Chao, K.R., Goodrich, J.K., Tiao, G., Lu, W., Riley-Gillis, B.M., Tsai, E.A., Kim, H.I., Zheng, X., et al. (2022). Systematic single-variant and gene-based association testing of thousands of phenotypes in 426,370 UK Biobank exomes. medRxiv, 2021.2006.2019.21259117. 10.1101/2021.06.19.21259117.
  20. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K., Aslibekyan, S., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet* 52, 969-983. 10.1038/s41588-020-0676-4.
  21. Curtis, D. (2022). Analysis of 200 000 exome-sequenced UK Biobank subjects illustrates the contribution of rare genetic variants to hyperlipidaemia. *J Med Genet* 59, 597-604. 10.1136/jmedgenet-2021-107752.
  22. Kurki, M.I., Karjalainen, J., Palta, P., Sipilä, T.P., Kristiansson, K., Donner, K., Reeve, M.P., Laivuori, H., Aavikko, M., Kaunisto, M.A., et al. (2022). FinnGen: Unique genetic insights from combining isolated population and national health register data. medRxiv, 2022.2003.2003.22271360. 10.1101/2022.03.03.22271360.
  23. Szustakowski, J.D., Balasubramanian, S., Kvikstad, E., Khalid, S., Bronson, P.G., Sasson, A., Wong, E., Liu, D., Wade Davis, J., Haefliger, C., et al. (2021). Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat Genet* 53, 942-948. 10.1038/s41588-021-00885-0.
  24. Van Hout, C.V., Tachmazidou, I., Backman, J.D., Hoffman, J.D., Liu, D., Pandey, A.K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* 586, 749-756. 10.1038/s41586-020-2853-0.
  25. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol* 17, 122. 10.1186/s13059-016-0974-4.
  26. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38, e164. 10.1093/nar/gkq603.
  27. Hebert, L.E., Weuve, J., Scherr, P.A., and Evans, D.A. (2013). Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology* 80, 1778-1783. 10.1212/WNL.0b013e31828726f5.
  28. Lee, S., Wu, M.C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics (Oxford, England)* 13, 762-775. 10.1093/biostatistics/kxs014.
  29. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83, 311-321. 10.1016/j.ajhg.2008.06.024.
  30. Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 86, 929-942. 10.1016/j.ajhg.2010.05.002.

31. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89, 82-93. 10.1016/j.ajhg.2011.05.029.
32. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* 95, 5-23. 10.1016/j.ajhg.2014.06.009.
33. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434-443. 10.1038/s41586-020-2308-7.
34. Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., Miranda, A., Fumis, L., Carvalho-Silva, D., Spitzer, M., et al. (2021). Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res* 49, D1302-D1310. 10.1093/nar/gkaa1027.



## Figure Legend

### **Figure 1. Manhattan plots displaying the $-\log_{10}(P)$ for the association of gene-based test results.**

Gene-based results showing the association strength between proxy AD and genes across the genome.

The dotted horizontal line represents the exome-wide association significance level,  $P < 2.5 \times 10^{-6}$ . To aid visualization, chromosomes are delimited by colors. Genomic positions are according to hg38.

### **Figure 2. Differential expression evidence from Agora.**

Differential expression results were queried from the Agora website. In the figure, the diameter and the color of the circles indicate p-value and log2 fold change, respectively.

Abbreviations: ACC, anterior cingulate cortex; CBE, cerebellum; DLPFL, dorsolateral prefrontal cortex; FP, frontal pole; IFG, inferior frontal gyrus; PCC, posterior cingulate cortex; PHG, parahippocampal gyrus; STG, superior temporal gyrus; TCX, temporal cortex.

**Table 1. Exome-wide significant gene-based results for proxy AD**

Chr	Pos	Gene	<i>P</i>	FinnGen $P_A/P_D^\dagger$
1	19074507-19214332	<i>UBR4</i>	3.84E-07	3.4E-5 <sup>†</sup>
1	56994777-58250547	<i>DABI</i>	1.34E-08	1.3E-5 <sup>†</sup>
1	58480718-58546726	<i>OMAI</i>	3.35E-09	1.3E-5 <sup>†</sup>
2	96335765-96377091	<i>NCAPH</i>	2.01E-06	1.6E-4 <sup>†</sup>
2	97113152-97264521	<i>ANKRD36</i>	6.01E-31	1.2E-4
2	168786538-168865339	<i>NOSTRIN</i>	5.73E-12	9.3E-5
2	168865044-168890443	<i>SPC25</i>	1.69E-09	1.2E-4
2	178431413-178451175	<i>PRKRA</i>	1.36E-07	3.8E-4
2	201288270-201364285	<i>FLACCI</i>	1.97E-06	-
2	219245468-219250337	<i>STK16</i>	4.88E-12	1.0E-4
2	219249709-219254608	<i>TUBA4A</i>	7.14E-12	1.0E-4
2	219253242-219272197	<i>TUBA4B</i>	7.97E-12	1.0E-4
3	50155057-50189075	<i>SEMA3F</i>	2.84E-07	8.7E-5
6	41158507-41163116	<i>TREM2</i>	6.89E-16	7.6E-5
6	45898450-46129819	<i>CLIC5</i>	2.82E-10	9.8E-4
6	80004146-80042651	<i>TTK</i>	2.10E-12	1.4E-5 <sup>†</sup>
7	1470276-1504389	<i>INTS1</i>	6.44E-09	3.2E-4
7	4892244-4959187	<i>MMD2</i>	2.80E-07	7.2E-5 <sup>†</sup>
7	100949533-100968347	<i>MUC3A</i>	9.61E-10	2.3E-4
7	142740234-142753076	<i>PRSS1</i>	1.34E-07	7.3E-4 <sup>†</sup>
8	86514421-86561498	<i>CPNE3</i>	6.75E-11	4.7E-4
8	86573607-86743675	<i>CNGB3</i>	7.33E-10	4.7E-4
9	33750465-33799231	<i>PRSS3</i>	9.70E-10	8.9E-4
9	3722372-33818868	<i>UBE2R2-AS1</i>	5.87E-10	1.2E-4 <sup>†</sup>
10	7158623-7411486	<i>SFMBT2</i>	2.00E-06	2.8E-4 <sup>†</sup>
10	19816214-20289856	<i>PLXDC2</i>	1.34E-07	2.7E-5
10	133262421-133276891	<i>ADAM8</i>	1.80E-09	2.3E-6
11	56700387-56701305	<i>OR9G1</i>	5.51E-09	6.4E-4
11	73950318-73970287	<i>DNAJB13</i>	8.01E-11	7.1E-5
11	102942994-102955732	<i>MMP13</i>	1.07E-06	2.5E-4 <sup>†</sup>
11	121452313-121633763	<i>SORL1</i>	5.78E-10	1.6E-5
12	11307076-11310436	<i>PRB4</i>	7.77E-15	6.2E-5 <sup>†</sup>
12	15322333-15598331	<i>PTPRO</i>	2.80E-07	3.4E-5
12	55835432-55842971	<i>MMP19</i>	2.81E-07	5.2E-4
14	32934784-33804176	<i>NPAS3</i>	1.55E-14	1.3E-4 <sup>†</sup>
14	35044906-35083383	<i>FAM177A1</i>	5.28E-07	6.5E-4

15	78959946-79090780	<b>RASGRF1</b>	2.92E-08	7.6E-4 <sup>†</sup>
15	80404381-80597933	<i>ARNT2</i>	2.25E-07	3.6E-4 <sup>†</sup>
17	38765690-38799905	<i>PIP4K2B</i>	3.13E-07	3.7E-4
17	43044294-43125364	<b>BRCA1</b>	3.26E-09	6.4E-4
17	44844280-44850480	<b>HIGD1B</b>	2.16E-10	6.1E-6
17	44849947-44899625	<b>EFTUD2</b>	2.15E-10	6.1E-6
17	75525079-75575209	<b>LLGL2</b>	1.91E-07	2.7E-4
19	1040106-1065572	<i>ABCA7</i>	4.71E-08	1.8E-7
19	8322583-8343262	<b>KANK3</b>	1.34E-07	2.5E-4
19	8848839-9010390	<b>MUC16</b>	8.73E-11	7.4E-4
19	44631171-44728395	<i>CEACAM16-AS1</i>	2.80E-07	9.3E-96
19	44643797-44666162	<i>PVR</i>	1.91E-06	4.4E-82
19	44671451-44684355	<i>CEACAM19</i>	6.31E-08	4.4E-82
19	44742619-44760044	<i>BCL3</i>	1.26E-15	8.9E-135
19	44758656-44758721	<b>MIR8085</b>	1.08E-15	-
19	44809058-44821421	<i>BCAM</i>	5.54E-29	0.0E+0
19	44846296-44889223	<i>NECTIN2</i>	2.68E-12	0.0E+0
19	44891219-44903689	<i>TOMM40</i>	5.37E-12	0.0E+0
19	44905795-44909395	<i>APOE</i>	5.23E-12	0.0E+0
19	44954584-44993346	<i>CLPTM1</i>	1.34E-08	0.0E+0
19	45039044-45070956	<i>CLASRP</i>	6.63E-20	7.6E-100
19	45071499-45076478	<b>ZNF296</b>	4.19E-14	8.9E-100
19	45085328-45090391	<i>GEMIN7-AS1</i>	1.02E-09	3.2E-76
19	45527426-45584802	<i>OPA3</i>	1.10E-06	3.4E-65
22	37805106-37807436	<b>HI-0</b>	2.07E-06	-
22	37807893-37817177	<b>GCAT</b>	1.94E-06	-
22	44668546-44737681	<b>PRR5</b>	7.67E-11	5.9E-4
22	44702197-44862784	<b>PRR5-ARHGAP8</b>	6.91E-11	5.7E-4 <sup>†</sup>

Genes meeting the exome-wide association significance,  $P < 2.5 \times 10^{-6}$ , for AD-by-proxy are presented.

Their corresponding association results for Alzheimer's disease in FinnGen (<https://www.finnngen.fi/>) are also shown (FinnGen  $P$ ,  $p$ -values  $< 1E-3$  are shown and the rest are displayed as a dash). Gene name is in boldface if it has not been previously reported for associations with Alzheimer's disease. Genomic positions are according to hg38. Chr, chromosome; Pos, position;  $P_A$ ,  $P$  value for Alzheimer's disease in FinnGen,  $P_D$ ,  $P$  value for dementia in Alzheimer disease in FinnGen.

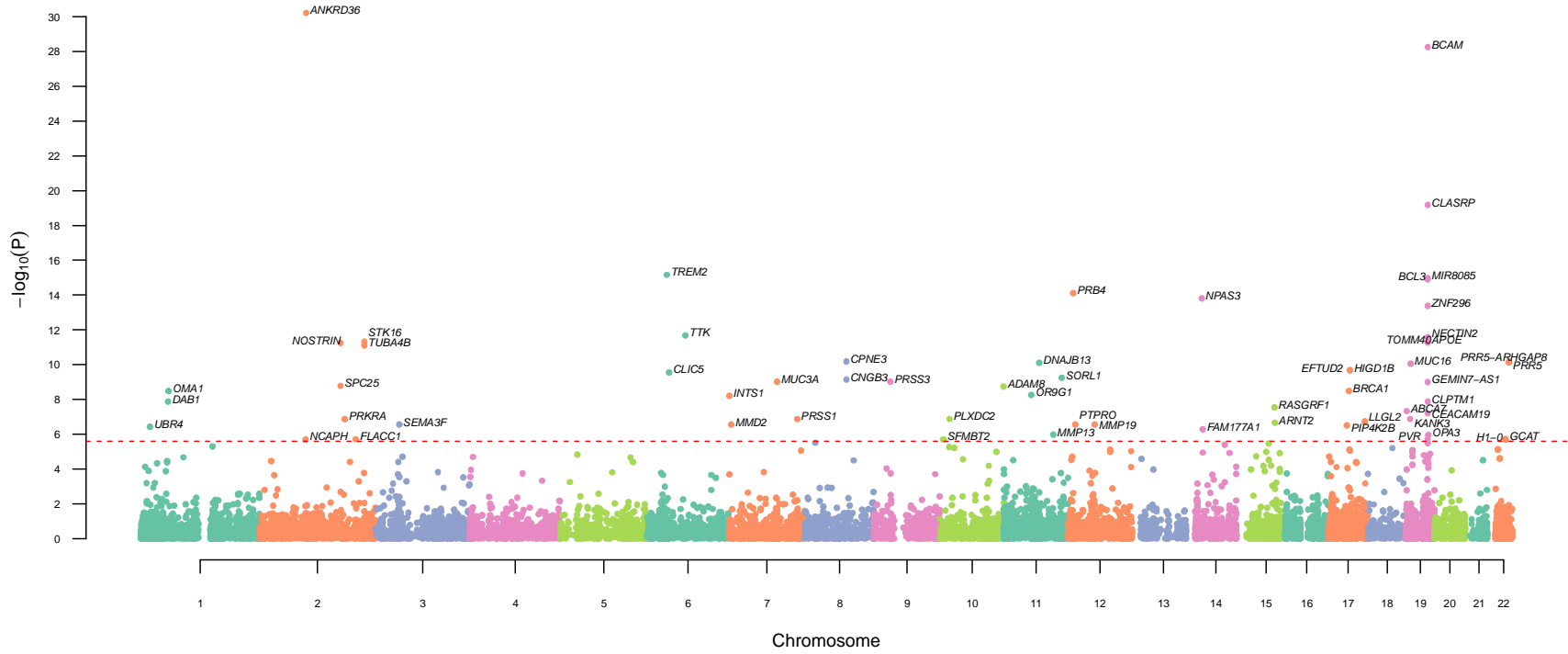
**Table 2. Known drug targets for the identified genes associated with Alzheimer’s**

Gene	Known Drugs	Mechanism of Action	Disease Information
<b><i>TUBA4A</i>*</b>	Davunetide	Tubulin stabiliser	Cognitive impairment Progressive supranuclear palsy Schizophrenia
	Sagopilone	Tubulin stabiliser	Central Nervous System Neoplasm Glioblastoma multiforme
	Lorvotuzumab mertansine	Tubulin inhibitor	Malignant peripheral nerve sheath tumor Glioblastoma multiforme
	ANG1005	Tubulin inhibitor	Malignant glioma Brain neoplasm Neoplasm
	Verubulin	Tubulin disrupting agent	Glioblastoma multiforme Brain neoplasm Neoplasm
	Lexibulin	Tubulin inhibitor	Glioblastoma multiforme Multiple myeloma
<b><i>MMP13</i>*</b>	Doxycycline	Matrix metalloproteinase 13 inhibitor	Alzheimer disease Brain injury
	Doxycycline Anhydrous	Matrix metalloproteinase 13 inhibitor	Alzheimer disease
<b><i>MUC16</i></b>	Oregovomab	Mucin-16 other	Ovarian cancer Pancreatic adenocarcinoma Peritoneum cancer Ovarian neoplasm Fallopian tube cancer Ovarian carcinoma
	Abagovomab	Mucin-16 other	Ovarian cancer Peritoneum cancer Fallopian tube cancer
	Sofituzumab Vedotin	Mucin-16 binding agent	Pancreatic carcinoma Ovarian cancer
<b><i>TTK</i></b>	BAY-1217389	Dual specificity protein kinase TTK inhibitor	Cancer
	BAY-1217389	Dual specificity protein kinase TTK inhibitor	Cancer

The table shows existing drugs that target the rare-variant genes identified.

\*These lists are not all inclusive. *TUBA4A* has a total of 19 drugs with 69 indications and *MMP13* has a total of 4 drugs with 103 indications.

**Figure 1. Manhattan plots displaying the  $-\log_{10}(P)$  for the association of gene-based test results.**



**Figure 2. Differential expression evidence from Agora.**

