HHS Public Access

Author manuscript

J Alzheimers Dis. Author manuscript; available in PMC 2024 June 06.

Published in final edited form as:

J Alzheimers Dis. 2023 ; 96(4): 1639–1649. doi:10.3233/JAD-230510.

Identifying Genes Associated with Alzheimer's Disease Using Gene-Based Polygenic Risk Score

Dongbing Lai[a,*], **Michael Zhang**[a], **Rudong Li**[a], **Chi Zhang**[a], **Pengyue Zhang**[b], **Yunlong Liu**[a], **Sujuan Gao**[b], **Tatiana Foroud**[a]

[a]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA

[b]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA

Abstract

Background: Except *APOE*, Alzheimer's disease (AD) associated genes identified in recent large-scale genome-wide association studies (GWAS) had small effects and explained a small portion of heritability. Many AD-associated genes have even smaller effects thereby sub-threshold *p*-values in large-scale GWAS and remain to be identified. For some AD-associated genes, drug targeting them may have limited efficacies due to their small effect sizes.

Objective: The purpose of this study is to identify AD-associated genes with sub-threshold *p*-values and prioritize drugs targeting AD-associated genes that have large efficacies.

Methods: We developed a gene-based polygenic risk score (PRS) to identify AD genes. It was calculated using SNPs located within genes and having the same directions of effects in different study cohorts to exclude cohort-specific findings and false positives. Gene co-expression modules and protein-protein interaction networks were used to identify AD-associated genes that interact with multiple other genes, as drugs targeting them have large efficacies via co-regulation or interactions.

Results: Gene-based PRS identified 389 genes with 164 of them not previously reported as AD-associated. These 389 genes explained 56.12%–97.46% SNP heritability; and they were enriched in brain tissues and 164 biological processes, most of which are related to AD and other neurodegenerative diseases. We prioritized 688 drugs targeting 64 genes that were in the same co-expression modules and/or PPI networks.

Conclusions: Gene-based PRS is a cost-effective way to identify AD-associated genes without substantially increasing the sample size. Co-expression modules and PPI networks can be used to identify drugs having large efficacies.

[*]Correspondence to: Dongbing Lai, PhD, 410 W. 10th Street, HS 4000, HITS, Indianapolis, IN 46202-3002, USA. Fax: +1 317 278 1100; dlai@iu.edu.

CONFLICTOFINTEREST

The authors declare no conflicts of interest.

SUPPLEMENTARYMATERIAL

The supplementary material is available in the electronic version of this article: https://dx.doi.org/10.3233/JAD-230510.

Keywords

Alzheimer's disease; gene-based polygenic risk score; gene-targeting drugs; gene co-expression module; protein-protein interaction network; SNP heritability

INTRODUCTION

Alzheimer's disease (AD) is the most common neurodegenerative disease and the most common cause of dementia. The estimated heritability of AD is between 60% and 80% [1]. Identifying genes contributing to the risk for AD will advance our knowledge of AD etiology and thus facilitate the development of novel therapeutic strategies, especially drugs targeting AD genes due to their known mechanisms [2–4]. Recent large-scale genome-wide association studies (GWAS) have identified >100 AD-associated genes; however, except *APOE*, all had small effects and together they only explained a small portion of AD heritability [5–15]; and many genes with even smaller effects remain to be discovered. Additionally, the small gene effects mean that drugs targeting some of them may have limited efficacies.

Increasing the AD GWAS sample size will identify more AD genes. However, this requires extensive and strategic data collection that cannot be achieved in near future. Furthermore, as demonstrated in a recent study, when the sample size reaches a certain level, further increasing it results in minimal returns in gene identification but results in dramatically increased cost [16]. In fact, as shown in a recent study with > 1M samples, the largest AD GWAS to date only identified 7 novel loci [15, 17]. While we should continue our effort to increase the sample size, novel methods to detect AD genes with sub-threshold *p*-values are urgently needed. Polygenic risk score (PRS) is the weighted sum of disease risk alleles and is used to predict disease risks. It requires discovery datasets to select SNPs and obtain their weights; then PRS is applied to target datasets that are independent of discovery datasets to predict disease risk. If PRS has high predictability (i.e., explained large portion of variation in the target datasets) and SNPs used in calculating PRS explain large portion of SNP heritability (h^2_{snp}), then these SNPs are likely disease-associated SNPs. Furthermore, if we also know the genes impacted by SNPs included in calculating PRS (i.e., gene-based PRS), then those genes are likely disease-associated genes. Therefore, gene-based PRS provides another way to identify disease genes with sub-threshold *p*-values.

To maximize PRS predictability, it is important to include more disease-associated SNPs and exclude unrelated SNPs (i.e., noise). SNPs used to calculate PRS are selected from the discovery dataset, which typically is a meta-analysis of multiple cohorts. Each cohort is usually ascertained using different strategies, sometimes even with different diagnostic methods. For instance, while cohorts from Kunkle et al. (2019) included clinically and neuropathologically defined AD cases and controls [8], the majority of AD cases from the UK Biobank (UKBB) were proxy AD cases, i.e., first-degree relatives of AD patients [12, 13, 18]. Therefore, some association findings may be limited only to a particular cohort (i.e., cohort-specific findings). Sophisticated meta-analysis algorithms can mitigate this problem but cannot eliminate cohort-specific findings, especially those with extremally

small *p*-values or those having low minor allele frequencies and only found in one or a few cohorts. In both cases, these findings result in *p*-values that remain small after meta-analysis. Additionally, due to the larg enumber of SNPs tested, many SNPs have small *p*-values simply due to random variations (i.e., false positives). These cohort-specific SNPs and false positives are indistinguishable from AD-associated SNPs with sub-threshold *p*-values and can be wrongly selected in calculating PRS, thereby reduce the PRS predictability. However, cohort-specific SNPs and false positives are less likely to have the same directions of effects and small *p*-values in all cohorts, i.e., they are not concordant; therefore, by keeping only concordant SNPs, the likelihood of including true disease-associated SNPs is dramatically increased and most noise are excluded, as a result, the PRS predictability can be substantially enhanced. We note that concordant SNPs can only be selected from cohorts with sufficiently large sample sizes as our goal is to detect SNPs with small effect sizes. For those SNPs, they may have opposite directions of effects in small cohorts due to random chance and/or cohort-specific effects, consequently, they will be excluded if only retaining concordant SNPs from all cohorts. In our previous work, PRS calculated using concordant SNPs had predictability comparable to the family history to predict the risk for alcohol use disorder [19]. To identify disease genes, we selected concordant SNPs located within gene boundaries and using this gene-based PRS, we identified 410 genes associated with alcohol use disorder [20].

In this study, we modified our gene-based PRS framework to identify AD-associated genes. We first used a leave-one-cohort-out strategy as described in Material and Methods section to identify concordant SNPs and AD-associated genes, then we calculated gene-based PRS by only using concordant SNPs in identified AD-associated genes. We tested whether these gene-based PRS had high predictability and identified AD-associated genes explained large portion of h^2_{snp} to ensure that they were truly AD related. We further confirmed the roles of identified genes in AD by searching previous GWAS of traits related to AD and other neurodegenerative diseases. Moreover, if these genes are truly AD related, then they should be expressed in AD-related tissues and presented in AD-related biological processes; therefore, we performed gene enrichment analyses to demonstrate this. One of the major goals to identify AD-associated genes is to search for drugs targeting them to facilitate the development of novel treatment methods. Since most genes have small effects, intuitively, drugs targeting them may have limited efficacies. However, for those genes in the same gene co-expression module or in the same protein-protein interaction (PPI) network, especially those genes interacting with multiple other genes (i.e., hub genes), drugs targeting them may have large efficacies as they may affect other genes via co-regulations and/or protein-protein interactions. Therefore, we identified gene co-expression modules and protein-protein interaction (PPI) networks, then searched the Drug Gene Interaction Database (DGIdb, v4.2.0) [21] for drugs targeting AD-associated genes. Those drugs targeting co-expressed genes and hub genes were prioritized.

MATERIALSANDMETHODS

Datasets

For PRS calculation, the discovery datasets are usually GWAS summary statistics, and the target datasets are typically individual level genotyping data if the goal is to predict an individual's disease risk. However, since our goal was to evaluate the overall predictability of gene-based PRS, therefore, GWAS summary statistics can also be used as the target dataset, and a specifically designed program, megaPRS as described below [22] was used for this purpose. Three independent, large-scale European ancestry AD GWAS summary statistics were used in our analyses: Kunkle et al. (21,982 cases and 41,944 controls) [8, 13], UKBB (53,042 cases and proxy cases, and 355,900 controls) [12, 13, 18], and FinnGen consortium (9,271 cases and 299,883 controls) [5,23]. UKBB is a relatively young population-based cohort (aged between 40–69, 898 AD cases) [24] and many high-risk individuals are too young to develop AD; therefore, proxy cases, defined as having at least one first-degree relative with AD/dementia, were also included (52,791 proxy cases); and those who reported not having AD and first-degree relatives with AD/dementia were considered as controls [12, 13, 18]. UKBB proxy cases were included in all recent large-scale AD GWAS except Kunkle et al. (2019) to dramatically increase the sample sizes [5–7, 13, 15]. Palindromic SNPs were excluded to avoid strand ambiguity. Only SNPs with rs numbers and the same reference and alternative alleles as in the 1000 Genomes Project (Phase 3, version5, NCBI GRCh37) were retained for analyses.

Leave-one-cohort-out strategy

We used the leave-one-cohort-out (LOCO) strategy to perform gene-based PRS analyses and to identify AD-associated genes (Fig. 1). Each large-scale AD GWAS was considered as a cohort. At each LOCO, two cohorts were used as the discovery datasets and the third cohort was used as the target dataset. This procedure was repeated three times until all cohorts had been used as the target dataset. Each LOCO identified a list of AD-associated genes, and we only retained those genes identified by all three LOCOs in subsequent analyses, as these genes were more likely to be true positives.

Identify AD-associated genes

During each LOCO, we first identified concordant SNPs, i.e., SNPs that had the same directions of effects and p-values < 0.05 (i.e., showing some degree of association) in both discovery datasets. Next, we identified candidate AD genes as those having at least one concordant SNP within the gene's boundaries. Gene boundaries were defined by using ANNOVAR (1kb from transcription start and end sites) [25] with gene information obtained from Ensembl (GRCh37, data accessed 2022–08-02). Due to the large amount of SNPs tested in GWAS, some genes irrelevant to AD may have one or a few concordant SNPs simply by chance, especially for those large genes; therefore, for all candidate AD genes, we further performed gene-based analysis by using MAGMA [26] to ensure that they were associated with AD at the gene level. MAGMA analyzes multiple SNPs in a gene simultaneously to evaluate their joint effect conditioned on gene size and density [26]; therefore, genes having concordant SNPs due to chance were excluded. MAGMA was performed on the meta-analysis results of both discovery datasets, and we retained genes

with MAGMA p-values < 0.05 (i.e., showing at least marginal association) as AD-associated genes.

Calculate PRS, evaluate PRS predictability, and estimate h^2_{snp} explained

In each LOCO, all concordant SNPs in AD-associated genes were included in the gene-based PRS calculation in the target dataset. The effect sizes of concordant SNPs used to calculate PRS were estimated from the meta-analysis of both discovery datasets by using METAL [27]. Due to the proxy case used in UKBB, small numbers of cases in both UKBB and FinnGen, as well as small portions of related individuals in these GWAS, we weighted the effect sizes using the inverse of standard errors instead of using sample sizes. We note that this was not optimal as using inverse variance resulted in biased effect size estimates but to a less extent than using sample sizes. Since *APOE* has a large effect on AD risk, we also performed gene-based PRS analyses by excluding *APOE* region (2 Mb from start and end of *APOE*). We chose a large region to exclude SNPs that are in linkage disequilibrium with *APOE* SNPs and SNPs that impact *APOE* long distantly.

We used megaPRS, which is a part of the genetic analysis software package LDAK [22] to evaluate PRS predictability. The prediction accuracy is measured by R^2. It is the squared correlation between the phenotype in the target GWAS summary statistics and the predicted phenotype based on the discovery dataset [22]. Higher R^2 indicates higher prediction accuracy. Since target datasets were also GWAS summary statistics; therefore, no additional covariate was included.

R^2 is a measure of variation due to genetic factors; therefore, it cannot exceed the estimated h^2_{snp} in the target dataset. The estimated h^2_{snp} may be low if the overall sample size or the number of AD cases is small, resulting even lower R^2. However, if genes identified were truly AD-associated, then h^2_{snp} explained by R^2 should be large. Therefore, we estimated h^2_{snp} in the target dataset using the BLD-LADK model implemented in LDAK [28] with h^2_{snp} estimated on the observed scale; then we calculated the percentage of h^2_{snp} explained by PRS as R^2/h^2_{snp}.

To demonstrate that the increase in the variance explained was not due to the megaPRS method used in this study, we also tested gene-based PRS by using individual-level genotype data. We used ADC7 (509 AD cases and 787 matched controls), which is a small dataset used as a replication cohort in Kunkle et al. (2019) [8]. For each discovery dataset, we calculated two types PRS: 1) Using all SNPs; 2) Using only those concordant SNPs located in AD-associated genes. PRS-CS was used to estimate the posterior effect size of each SNP [29]. PLINK [30,31] was used to calculate PRS, with sex, age, the number of *APOE* E4 alleles, and the first 10 principal components of genetic ancestry included as covariates. For each analysis, we calculated the delta R^2, which is the difference between the variances explained by fitting models with and without PRS, respectively.

For both PRS analysis and h^2_{snp} estimation, European ancestry samples from the 1000 Genomes Project was used as the reference panel to prune SNPs that are in linkage disequilibrium with each other following the steps recommended by LDAK [28]. PRS

analysis and h^2_{snp} estimation were performed at each LOCO, with and without *APOE* region.

Search GWAS catalog and gene enrichment analyses

For those genes identified in all three LOCOs, to confirm their roles in AD, we first searched the GWAS catalog [32] using FUMA (Functional Mapping and Annotation of Genome-wide association studies) [33] to identify genes previously implicated in GWAS of traits related to AD and other neurodegenerative diseases such as aging, amyloid-beta measurement, cerebellar volume measurement, cognitive function, dementia, memory, p-tau measurement, etc. We then performed two types of gene enrichment analyses. The first one was tissue specificity enrichment analysis using differentially expressed gene (DEG) sets, which were pre-calculated by FUMA using gene expression data from 54 tissue types obtained from GTEx (v8) [34]. The second approach was gene ontology enrichment analysis using PANTHER [35] implemented in the Gene Ontology (GO) Resource (http://geneontology.org/, accessed: 2022–09-15). We focused on GO Biological Processes (GOBPs), which were compiled by using multiple molecular activities. For both enrichment analyses, Benjamini-Hochberg false discovery rate (FDR) [36]< 0.05 was considered as significant.

For those identified AD-associated genes that were not previously reported as AD related, to further demonstrate that they were associated with AD, we also checked the minimum SNP *P*-values in these genes and gene-based *p*-values in Kunkle et al. (2019), UKBB, and FinnGen. We also performed gene enrichment analyses by using these genes only.

Prioritize drugs targeting AD-associated genes by using gene co-expression modules and protein-protein interaction networks

RNA sequencing data from human cortex ($N = 255$) were obtained from GTEx (V8) [34]. Genes with more than half of samples with read counts < 10 were excluded. Transcript per million (TPM) calculated from the read count was used. Local maximal Quasi-Clique Merger (lmQCM) [37] implemented in the network analysis tool suite TSUNAMI [38] was used to detect co-expression modules, taking advantage of its ability to identify small densely connected co-expression modules.

For proteins encoded by identified AD associated genes, we used the STRING database (https://string-db.org/, accessed: 2022–09-18) [39] to identify PPI networks. STRING database includes known and predicted PPI networks, which are derived from high-throughput lab experiments, previous knowledge in databases, automated text mining, genomic context predictions, and conserved Co-expression [39]. We only retained high confidence interactions from human (i.e., having confidence score ≥ 0.7 or obtained from high-throughput laboratory experiments and previous knowledge in databases).

We searched the Drug Gene Interaction Database (DGIdb, v4.2.0) [21] for drugs targeting AD associated genes identified in all three LOCOs. Only drugs with the Anatomical Therapeutic Chemical (ATC) codes (obtained from the Kyoto Encyclopedia of Genes and Genomics (KEGG): https://www.genome.jp/kegg/drug/) were retained. Drug targeting genes

in the same co-expression modules or PPI networks, or targeting genes that interact with ≥ 3 other genes (i.e., hub genes) in the same PPI network were prioritized.

RESULTS

Identification of AD-associated genes, variation explained by PRS, and h^2_{snp} explained

Table 1 summarized the gene-based PRS analyses and h^2_{snp} for each LOCO. Numbers of identified AD-associated genes ranged from 1,711 to 2,003 and about 400K concordant SNPs were included in each LOCO. Variation explained (R^2) ranged from 0.08 to 0.23 and explained more than half to almost all h^2_{snp} (56.12%–97.46%). After excluding *APOE*, R^2 ranged from 0.02 to 0.06 and these genes explained 29.93%–45.79% of h^2_{snp}. There were 389 AD-associated genes identified in all three LOCOs (Supplementary Table 1, with MAGMA *p*-values). For gene-based PRS analyses using ADC7, PRS were significant when Kunkle et al. (2019) and FinnGen were used as the discovery datasets. Using concordant SNPs in AD-associated genes explained more variance in the data (delta R^2 = 0.72%, p = 0.001) than using all SNPs (delta R^2 = 0.29%, p = 0.04), indicating that the increased variance explained was not due to the megaPRS method.

Searching GWAS catalog, gene enrichment analyses

More than half of the genes (225) were reported in GWAS of traits related to AD and other neurodegenerative diseases (Supplementary Table 2); the remaining (164 genes) were not reported previously (Supplementary Table 2). The minimum SNP *p*-values and gene *p*-values for those 164 genes from Kunkle et al, UKBB, and FinnGen are shown in Supplementary Table 3. All 164 genes have minimum SNP *p*-values < 0.05 and thus showed some degree of associations. About half these genes did not have any gene level associations (*p*-values ≥ 0.05) in each dataset but they had *p*-values < 0.05 in meta-analyses (Supplementary Table 1) by using concordant SNPs only, indicating that excluding most study-specific findings and false positives in gene level analysis resulted in more findings.

Tissue type gene enrichment analysis results are shown in Supplementary Figure 1 and Supplementary Table 4. AD-associated genes were significantly (FDR *p*-value < 0.05) enriched in 20, 43, and 41 upregulated, downregulated, and both up- and downregulated tissue-specific DEG sets, respectively, including all 13 brain tissues, adipose visceral omentum, aorta, spleen, and whole blood, etc. The top 10 most enriched tissues are all from brain, indicating that identified genes are likely AD related. GOBPs enrichment analysis results are shown in Supplementary Table 5 and 164 GOBPs had FDR *p*-values < 0.05, including those related to AD and other neurodegenerative diseases such as regulation of neurofibrillary tangle assembly, chylomicron remnant clearance, regulation of tau-protein kinase activity, etc. This again indicated that identified genes are likely AD related. For those 164 genes not previously reported as AD related, they were also majorly enriched in brain tissues, and they are likely AD related. However, they were not significantly enriched in any GOBPs. This is not a surprise as these GOBP were derived based on known knowledges of AD and these genes were not reported previously.

Using gene co-expression modules and protein-protein interaction networks to prioritize drugs targeting AD-associated genes

There were 348 genes that had gene expression readcount ≥ 10 for more than half of the GTEx cortex samples. Four co-expression modules were identified with 58, 58, 18, and 15 genes in modules 1, 2, 3, and 4, respectively (Supplementary Table 6). Proteins encoded by 98 AD-associated genes interacted with each other directly or via proteins encoded by other genes (connecting genes) and they formed 15 PPI networks (Supplementary Figure 3 and Supplementary Table 7). There were 54 genes (1 connecting genes; well-established AD genes such as *APOE* and *APP* were in this network), 15 genes (9 connecting genes), 11 genes (1 connecting genes), and 8 genes (1 connecting genes) in networks 1 to 4, respectively. All other networks had 3 or 2 genes. There were 37 hub genes in these 15 PPI networks.

We identified 878 drugs targeting 102 AD-associated genes from DGIdb (Supplementary Table 8). In total, 688 drugs targeted 64 genes (24 of them were hub genes) that were in the same co-expression modules and/or PPI networks.

DISCUSSION

In this study, using gene-based PRS, we identified 389 AD-associated genes. More than half of them were reported in previous GWAS of AD or other neurodegenerative diseases related traits; others showed some degree of associations in recent large-scale AD GWAS but were not genome-wide significant. In our study, these genes together explained a large portion of AD h^2_{snp}. These AD-associated genes were enriched in brain tissue and biological processes related to AD. We also prioritized drugs targeting genes in the same co-expression modules and/or PPI networks for future investigation of AD treatment.

We used concordant SNPs located within gene boundaries to calculate gene-based PRS and to identify AD genes with sub-threshold *p*-values. Based on the variation explained by PRS and the estimated h^2_{snp}, MAGMA gene-based *p*-values (< 0.05 in all three LOCOs), as well as tissues and biological processes that they were enriched in, the identified AD-associated genes with sub-threshold *p*-values are likely to be AD-associated and they should be prioritized in future functional studies to validate their roles in AD risk. This simple strategy is much more cost-effective than recruiting large samples to have sufficient statistical power to identify these genes using genome-wide levels of significance, or functionally testing all possible AD genes showing some degrees of association in large-scale GWAS. We acknowledge that some true AD-associated genes may be missed by this strategy. However, those genes would not have concordant SNPs within the gene boundaries and/or have MAGMA gene-based *p*-value ≥ 0.05; therefore, even if they are true AD-associations, they may have minimal or negligible effects on AD risk. We limited to genes with concordant SNPs within gene boundaries and acknowledge that many regulatory SNPs, especially those identified as eQTLs, are located out of gene boundaries. However, as shown in a recent study, only up to 8% of eQTLs are disease relevant [40]; therefore, including eQTLs would likely introduce more noise, resulting in low predictability of PRS.

The estimated heritability of AD from twin studies [1] is 60% to 80% but h^2_{snp} estimates from recent large-scale GWAS were ≤ 9% [5, 6, 8, 11, 15, 17, 18], aphenomenon called the missing heritability problem [41, 42]. Furthermore, as noted by a recent study, paradoxically with substantially increased sample sizes, although more genome-wide significant SNPs were identified, the h^2_{snp} estimated actually became smaller [17]. One theory about the missing heritability is that many disease-causing variants such as rare SNPs or CNVs cannot be detected by GWAS [41, 42] but this cannot explain why h^2_{snp} estimates became smaller with larger and more powerful GWAS. Our study demonstrated that the missing heritability was at least partially due to including erroneous SNPs in estimating h^2_{snp}. For instance, the reported h^2_{snp} in Kunkle et al. (2019) was only 7% using the LDsc model [8]. Although we re-estimated h^2_{snp} using the same BLD-LDAK model as used in this study and it increased to 21%, however, by focusing on concordant SNPs in identified AD-associated genes, the h^2_{snp} estimation increased to 40%, which is about 1/2 to 2/3 of the heritability estimated from twin studies [1]. This indicated that: 1) many AD-associated SNPs were already identified by large-scale GWAS but with sub-threshold *p*-values and thus indistinguishable from cohort-specific findings and false positives; 2) Including all SNPs in calculating h^2_{snp} resulted in the inclusion of both cohort-specific findings and false positive SNPs thereby dramatically underestimating the h^2_{snp}. Increasing AD GWAS sample sizes by adding additional large cohorts will generate more cohort-specific findings and false positives, and this explains the smaller h^2_{snp} with the increasing sample sizes, as shown in a recent study [43]. Therefore, we recommend that for those SNPs that are not concordant among different cohorts, as long as each cohort has a reasonable large sample size (e.g., tens of thousands of samples, preferably clinically determined cases and well-matched controls) to warrant sufficient statistical power, they should be removed from h^2_{snp} estimation and PRS analysis to increase the likelihood of selecting true disease-associated SNPs and excluding noise. In our study, the estimated h^2_{snp} using concordant SNPs in Kunkle et al. (2019) was still smaller than the heritability estimated from twin studies. However, among 1,745 genes (Table 1) included in estimating h^2_{snp}, only 389 genes were common in all three LOCOs, indicating that many of them may still be cohort-specific and/or false positives. Additionally, we focused on concordant SNPs within gene boundaries and excluded regulatory SNPs located outside of genes; therefore, our smaller h^2_{snp} estimate as compared to twin studies is actually expected. Lastly, as in Kunkle et al. (2019) [8], we reported h^2_{snp} on the observed scale, which is study-specific and cannot be used to compare with other studies. Ideally, h^2_{snp} estimated on the liability scale should be used; however, it requires accurate estimation of AD prevalence, which is age dependent. Different cohorts have different age distributions, making it challenging to precisely estimate AD prevalence, consequently, h^2_{snp} estimation on the liability scale may be biased and thus was not used in this study. In the analyses of excluding *APOE* region, the estimated h^2_{snp} and h^2_{snp} explained were about half of those with *APOE* in all three LOCOs. Such a reduction is expected due to the large effect in *APOE* region.

For UKBB and FinnGen, although the estimated h^2_{snp} were more than doubled those from the original report [17], they were still only around 8% in our study.As summarized by a recent study, this is largely due to the use of proxy cases in UKBB and population controls in both UKBB and FinnGen [17]. In UKBB, there were 898 case and 52,791 proxy cases

(98.33% of the entire cases) [13]. Using proxy cases reduced the effective sample size by a factor of up to 4.9 and limited the ability to detect SNPs with small effects [12]. Furthermore, proxy case was defined as having a parent or sibling with AD or dementia [13, 18], therefore, some of the proxy cases will never develop AD, resulting in underestimated SNP effect sizes [7]. Both UKBB and FinnGen used population controls instead of clinically screened and age matched controls as used in Kunkle et al. (2019) [8, 12, 13, 17, 18]. Controls were younger than cases and simply adjusting age in analyses would result in spurious findings due to collider bias because age is also a significant contributor to AD and is heritable [44]. Furthermore, since AD is age related and genetic factors are more likely to determine when to develop AD instead of whether to develop AD [45, 46], therefore, heritability estimate is also age dependent [17]. All of these resulted in lower statistical power in UKBB and FinnGen, hence the small estimate of h^2_{snp}, despite the large overall sample sizes. Our PRS analysis results also demonstrated the lower statistical power of UKBB and FinnGen. When Kunkle et al, 2019 was included as the discovery dataset, the h^2_{snp} explained were 97.46% and 69.53% for UKBB and FinnGen (*APOE* included), respectively; however, the h^2_{snp} explained when using UKBB and FinnGen as the discovery datasets was only 56.12% (Table 1), indicating that at least the SNP effect sizes were underestimated. While biobanks provide unprecedented opportunity to quickly increase the sample size, more effort should be devoted to accurately define cases and controls as the statistical power may only increase minimally even with the dramatically increased sample size.

Although small gene effects in GWAS do not necessarily mean small efficacies of drugs targeting these genes, methods to prioritize the large number of drugs identified are necessary. We used both co-expression modules and PPI networks as gene and protein expressions are only weakly correlated [47–50] and many drugs were designed to work on proteins. Among prioritized drug-targeting genes, *PSMB5* and *PSMC6* were in the same co-expression module and PPI network, and both interacting with 9 other genes (majority of them were connecting genes). Both genes and most of their interacting genes belong to proteasomes and regulate proteins degradation; not surprisingly, proteasomes inhibitors such as carfilzomib and bortezomib target these genes. These drugs were designed to treat cancer [21] but may provide some clues for treating AD. In addition, genes in PPI network 1 and drugs targeting them were of particular interest. Among them, 12 genes (*APOE*, *APP*, *CLU*, *EGFR*, *ERBB4*, *HBEGF*, *HLA-DQA1*, *HLA-DRA*, *HLA-DRB1*, *MS4A2*, *PLCG2*, and *PTK2B*) were hub genes and were also genome-wide significant in previous AD GWAS. Aducanumab targeting *APP* has been approved by FDA to treat AD [21]. Other drugs targeting these genes such as atorvastatin (anticholesterolaemic), bevacizumab (antineoplastic), cetuximab (antineoplastic), floxacillin (antibacterials), fluvastatin (anticholesteremic), irbesartan (antihypertensive), lapatinib (antineoplastic), lorazepam (antianxiety), lubiprostone (anti-constipation), omalizumab (antiasthmatic), paliperidone, (antipsychotic), and prednisone (anti-inflammatory), have also been approved by FDA [21] but could be potentially repurposed to treat AD. Drugs or combinations of them targeting aforementioned genes should have the top priority for future investigations. However, we note that while drugs targeting other genes were not prioritized, they should not be ignored. For example, *ACE* was not identified in any co-expression

module and PPI network but one of the drugs targeting it, germfibrozil, which is approved by FDA to treat hyperlipidemias, was also identified in a recent study through a deep learning framework as potential drug to treat AD [51]. For the drugs identified in our analyses, the next step is to use real-world patient data to investigate their efficacies. For example, testing whether those taking these drugs have different prevalence of AD compared to those not taking the drugs with the same health conditions. Another critical part for drug development is side effects. While side effects of those drugs are available from previous clinical trials; however, those trials are not designed to treat AD. Side effects can also be investigated by using real-world patient data by focusing on AD patients only.

This study has several limitations. First, we used concordant SNPs located within gene boundaries to define candidate AD genes. While in general this is the most conservative way to define disease genes, it misses genes that their causal SNPs are outside gene boundaries, e.g., those SNPs acting long distantly through chromatin interactions. Second, if an AD-associated gene overlaps with other genes that are not AD-associated, and concordant SNPs happen to be in the overlapped regions, then all genes would be selected as AD-associated. Third, we used GWAS summary statistics in this study due to the lack of individual level genotype data. GWAS summary statistics only has aggregated information and this may cause less accurate PRS predictability estimation. Fourth, due to limited sample sizes, under-represented populations such as those with African and Hispanic ancestries were not studied. Fifth, for some SNPs, the allele frequencies are different between FinnGen and other samples even they are all European ancestries.

In summary, we have shown that gene-based PRS can be used to identify AD-associated genes with sub-threshold *p*-values using current available data. These genes will be validated in functional studies in the next step. We also prioritized drugs targeting identified AD genes and will investigate their potentials to treat AD in future studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We want to acknowledge the participants and investigators of the FinnGen study.

The authors acknowledge the Indiana University Pervasive Technology Institute for providing [HPC (Big Red II, Karst, Carbonate), visualization, database, storage, or consulting] resources that have contributed to the research results reported within this paper.

FUNDING

D. Lai and T. Foroud are supported by NIH grant U24AG021886.

S. Gao is supported by NIH grant P30072976.

DATAAVAILABILITY

GWAS summary statistics included in this study can be downloaded here: https://www.ebi.ac.uk/gwas/search?query=GCST90012877 (Kunkle et al. (2019) and UKBB) and https://www.finngen.fi/en/access results (FinnGen consortium).

REFERENCES

[1]. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A, Pedersen NL (2006) Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry 63, 168–174. [PubMed: 16461860]

[2]. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, Vitsios D, Deevi SVV, Mackay A, Muthas D, Hühn M, Monkley S, Olsson H; AstraZeneca Genomics Initiative; Wasilewski S, Smith KR, March R, Platt A, Haefliger C, Petrovski S (2021) Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature 597, 527–532. [PubMed: 34375979]

[3]. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, Whittaker JC, Sanseau P (2015) The support of human genetic evidence for approved drug indications. Nat Genet 47, 856–860. [PubMed: 26121088]

[4]. King EA, Davis JW, Degner JF (2019) Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genet 15, e1008489. [PubMed: 31830040]

[5]. Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, Naj AC, 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. [PubMed: 35379992]

[6]. de Rojas I, Moreno-Grau S, Tesi N, Grenier-Boley B, Andrade V, Jansen IE, Pedersen NL, Stringa N, Zettergren A, Hernández I, et al. (2021) Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. Nat Commun 12, 3417. [PubMed: 34099642]

[7]. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, Sealock J, Karlsson IK, Hägg S, Athanasiu L, Voyle N, Proitsi P, Witoelar A, Stringer S, Aarsland D, Almdahl IS, Andersen F, Bergh S, Bettella F, Bjornsson S, Brækhus A, Bråthen G, de Leeuw C, Desikan RS, Djurovic S, Dumitrescu L, Fladby T, Hohman TJ, Jonsson PV, Kiddle SJ, Rongve A, Saltvedt I, Sando SB, Selbæk G, Shoai M, Skene NG, Snaedal J, Stordal E, Ulstein ID, Wang Y, White LR, Hardy J, Hjerling-Leffler J, Sullivan PF, van der Flier WM, Dobson R, Davis LK, Stefansson H, Stefansson K, Pedersen NL, Ripke S, Andreassen OA, Posthuma D (2019) Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet 51, 404–413. [PubMed: 30617256]

[8]. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, Boland A, Vronskaya M, van der Lee SJ, Amlie-Wolf A, et al. (2019) 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. [PubMed: 30820047]

[9]. Kunkle BW, Schmidt M, Klein HU, Naj AC, Hamilton-Nelson KL, Larson EB, Evans DA, De Jager PL, Crane PK, Buxbaum JD, Ertekin-Taner N, Barnes LL, Fallin MD, Manly JJ, Go RCP, Obisesan TO, Kamboh MI, Bennett DA, Hall KS, Goate AM, Foroud TM, Martin ER, Wang LS, Byrd GS, Farrer LA, Haines JL, Schellenberg GD, Mayeux R, Pericak-Vance MA, Reitz C, Graff-Radford NR, Martinez I, Ayodele T, Logue MW, Cantwell LB, Jean-Francois M, Kuzma AB, Adams LD, Vance JM, Cuccaro ML, Chung J, Mez J, Lunetta KL, Jun GR, Lopez OL, Hendrie HC, Reiman EM, Kowall NW, Leverenz JB, Small SA, Levey AI, Golde TE, Saykin AJ, Starks TD, Albert MS, Hyman BT, Petersen RC, Sano M, Wisniewski T, Vassar R, Kaye JA, Henderson VW, DeCarli C, LaFerla FM, Brewer JB, Miller BL, Swerdlow RH, Van Eldik LJ, Paulson HL, Trojanowski JQ, Chui HC, Rosenberg RN, Craft S, Grabowski TJ, Asthana S, Morris JC, Strittmatter SM, Kukull WA (2021) Novel Alzheimer disease risk loci and pathways in African American individuals using the African Genome Resources Panel: a meta-analysis. JAMA Neurol 78, 102–113. [PubMed: 33074286]

[10]. Lake J, Warly Solsberg C, Kim JJ, Acosta-Uribe J, Makarious MB, Li Z, Levine K, Heutink P, Alvarado C, Vitale D, Kang S, Gim J, Lee K, Pina-Escudero SD, Ferrucci L, Singleton AB, Blauwendraat C, Nalls MA, Yokoyama JS, Leonard HL (2022) Multi-ancestry meta-analysis and fine-mapping in Alzheimer's Disease. medRxiv, 2022.2008.2004.22278442.

[11]. Lambert J-C, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, Jun G, DeStefano AL, Bis JC, Beecham GW, et al. (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 45, 1452–1458. [PubMed: 24162737]

[12]. Liu JZ, Erlich Y, Pickrell JK (2017) Case-control association mapping by proxy using family history of disease. Nat Genet 49, 325–331. [PubMed: 28092683]

[13]. Schwartzentruber J, Cooper S, Liu JZ, Barrio-Hernandez I, Bello E, Kumasaka N, Young AMH, Franklin RJM, Johnson T, Estrada K, Gaffney DJ, Beltrao P, Bassett A (2021) Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. Nat Genet 53, 392–402. [PubMed: 33589840]

[14]. Sherva R, Zhang R, Sahelijo N, Jun G, Anglin T, Chanfreau C, Cho K, Fonda JR, Gaziano JM, Harrington KM, Ho Y-L, Kremen W, Litkowski E, Lynch J, Neale Z, Roussos P, Marra D, Mez J, Miller MW, Salat DH, Tsuang D, Wolf E, Zeng Q, Panizzon M, Merritt VC, Farrer LA, Hauger RL, Logue MW (2022) African Ancestry GWAS of dementia in a large military cohort identifies significant risk loci. medRxiv, 2022.2005.2025.22275553.

[15]. Wightman DP, Jansen IE, Savage JE, Shadrin AA, Bahrami S, Holland D, Rongve A, Børte S, Winsvold BS, Drange OK, Martinsen AE, Skogholt AH, Willer C, Bråthen G, Bosnes I, Nielsen JB, Fritsche LG, Thomas LF, Pedersen LM, Gabrielsen ME, Johnsen MB, Meisingset TW, Zhou W, Proitsi P, Hodges A, Dobson R, Velayudhan L, Heilbron K, Auton A; 23andMe Research Team; Sealock JM, Davis LK, Pedersen NL, Reynolds CA, Karlsson IK, Magnusson S, Stefansson H, Thordardottir S, Jonsson PV, Snaedal J, Zettergren A, Skoog I, Kern S, Waern M, Zetterberg H, Blennow K, Stordal E, Hveem K, Zwart JA, Athanasiu L, Selnes P, Saltvedt I, Sando SB, Ulstein I, Djurovic S, Fladby T, Aarsland D, Selbæk G, Ripke S, Stefansson K, Andreassen OA, Posthuma D (2021) A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nat Genet 53, 1276–1282. [PubMed: 34493870]

[16]. Sullivan PF, Agrawal A, Bulik CM, Andreassen OA, Borglum AD, Breen G, Cichon S, Edenberg HJ, Faraone SV, Gelernter J, Mathews CA, Nievergelt CM, Smoller JW, O'Donovan MC (2018) Psychiatric genomics: an update and an agenda. Am J Psychiatry 175, 15–27. [PubMed: 28969442]

[17]. Escott-Price V, Hardy J (2022) Genome-wide association studies for Alzheimer's disease: bigger is not always better. Brain Commun 4, fcac125. [PubMed: 35663382]

[18]. Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, Davies G, Ritchie CW, Gale CR, Starr JM, Goate AM, Porteous DJ, Yang J, Evans KL, Deary IJ, Wray NR, Visscher PM (2018) GWAS on family history of Alzheimer's disease. Transl Psychiatry 8, 99. [PubMed: 29777097]

[19]. Lai D, Johnson EC, Colbert S, Pandey G, Chan G, Bauer L, Francis MW, Hesselbrock V, Kamarajan C, Kramer J, Kuang W, Kuo S, Kuperman S, Liu Y, McCutcheon V, Pang Z, Plawecki MH, Schuckit M, Tischfield J, Wetherill L, Zang Y, Edenberg HJ, Porjesz B, Agrawal A, Foroud T (2022) Evaluating risk for alcohol use disorder: Polygenic risk scores and family history. Alcohol Clin Exp Res 46, 374–383. [PubMed: 35267208]

[20]. Lai D, Schwantes-An TH, Abreu M, Chan G, Hesselbrock V, Kamarajan C, Liu Y, Meyers JL, Nurnberger JI Jr., Plawecki MH, Wetherill L, Schuckit M, Zhang P, Edenberg HJ, Porjesz B, Agrawal A, Foroud T (2022) Gene-based polygenic risk scores analysis of alcohol use disorder in African Americans. Transl Psychiatry 12, 266. [PubMed: 35790736]

[21]. Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, Griffith M, Griffith OL, Wagner AH (2021) Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Res 49, D1144–D1151. [PubMed: 33237278]

[22]. Zhang Q, Privé F, Vilhjálmsson B, Speed D (2021) Improved genetic prediction of complex traits from individual-level data or summary statistics. Nat Commun 12, 4192. [PubMed: 34234142]

[23]. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, Reeve MP, Laivuori H, Aavikko M, Kaunisto MA, et al. (2023) FinnGen provides genetic insights from a well-phenotyped isolated population. Nature 613, 508–518. [PubMed: 36653562]

[24]. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J (2018) The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209. [PubMed: 30305743]

[25]. Wang K, Li M, Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164–e164. [PubMed: 20601685]

[26]. de Leeuw CA, Mooij JM, Heskes T, Posthuma D (2015) MAGMA: Generalized gene-set analysis of GWAS data. PLoS Comput Biol 11, e1004219. [PubMed: 25885710]

[27]. Willer CJ, Li Y, Abecasis GR (2010) METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191. [PubMed: 20616382]

[28]. Speed D, Balding DJ (2019) SumHer better estimates the SNP heritability of complex traits from summary statistics. Nat Genet 51, 277–284. [PubMed: 30510236]

[29]. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW (2019) Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun 10, 1776. [PubMed: 30992449]

[30]. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7. [PubMed: 25722852]

[31]. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559–575. [PubMed: 17701901]

[32]. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Flicek P, Burdett T, Hindorff LA, Cunningham F, Parkinson H (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47, D1005–D1012. [PubMed: 30445434]

[33]. Watanabe K, Taskesen E, van Bochoven A, Posthuma D (2017) Functional mapping and annotation of genetic associations with FUMA. Nat Commun 8, 1826. [PubMed: 29184056]

[34]. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, Berghuis B, Turner L, Hudson E, Feenstra K, Sobin L, Robb J, Branton P, Korzeniewski G, Shive C, Tabor D, Qi L, Groch K, Nampally S, Buia S, Zimmerman A, Smith A, Burges R, Robinson K, Valentino K, Bradbury D, Cosentino M, Diaz-Mayoral N, Kennedy M, Engel T, Williams P, Erickson K, Ardlie K, Winckler W, Getz G, DeLuca D, MacArthur D, Kellis M, Thomson A, Young T, Gelfand E, Donovan M, Meng Y, Grant G, Mash D, Marcus Y, Basile M, Liu J, Zhu J, Tu Z, Cox NJ, Nicolae DL, Gamazon ER, Im HK, Konkashbaev A, Pritchard J, Stevens M, Flutre T, Wen X, Dermitzakis ET, Lappalainen T, Guigo R, Monlong J, Sammeth M, Koller D, Battle A, Mostafavi S, McCarthy M, Rivas M, Maller J, Rusyn I, Nobel A, Wright F, Shabalin A, Feolo M, Sharopova N, Sturcke A, Paschal J, Anderson JM, Wilder EL, Derr LK, Green ED, Struewing JP, Temple G, Volpi S, Boyer JT, Thomson EJ, Guyer MS, Ng C, Abdallah A, Colantuoni D, Insel TR, Koester SE, Little AR, Bender PK, Lehner T, Yao Y, Compton CC, Vaught JB, Sawyer S, Lockhart NC, Demchok J, Moore HF (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580–585. [PubMed: 23715323]

[35]. Mi H, Muruganujan A, Casagrande JT, Thomas PD (2013) Large-scale gene function analysis with the PANTHER classification system. Nat Protoc 8, 1551–1566. [PubMed: 23868073]

[36]. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. J R Stat Soc B Stat Methodol 57, 289–300.

[37]. Zhang J, Huang K (2014) Normalized lmQCM: an algorithm for detecting weak quasi-cliques in weighted graph with applications in gene co-expression module discovery in cancers. Cancer Inform 13, 137–146. [PubMed: 27486298]

[38]. Huang Z, Han Z, Wang T, Shao W, Xiang S, Salama P, Rizkalla M, Huang K, Zhang J (2021) TSUNAMI: translational bioinformatics tool suite for network analysis and mining. Genomics Proteomics Bioinformatics 19, 1023–1031. [PubMed: 33705981]

[39]. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 49, D605–d612. [PubMed: 33237311]

[40]. Connally NJ, Nazeen S, Lee D, Shi H, Stamatoyannopoulos J, Chun S, Cotsapas C, Cassa CA, Sunyaev SR (2022) The missing link between genetic association and regulatory function. Elife 11, e74970. [PubMed: 36515579]

[41]. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456, 18–21. [PubMed: 18987709]

[42]. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461, 747–753. [PubMed: 19812666]

[43]. Wang X, Walker A, Revez JA, Ni G, Adams MJ, McIntosh AM, Wray NR, Ripke S, Mattheisen M, Trzaskowski M, et al. (2023) Polygenic risk prediction: why and when out-of-sample prediction R2 can exceed SNP-based heritability. Am J Hum Genet 110, 1207–1215. [PubMed: 37379836]

[44]. Day FR, Loh PR, Scott RA, Ong KK, Perry JR (2016) A robust example of collider bias in a genetic association study. Am J Hum Genet 98, 392–393. [PubMed: 26849114]

[45]. Meyer MR, Tschanz JT, Norton MC, Welsh-Bohmer KA, Steffens DC, Wyse BW, Breitner JC (1998) APOE genotype predicts when–not whether–one is predisposed to develop Alzheimer disease. Nat Genet 19, 321–322. [PubMed: 9697689]

[46]. Reiman EM, Arboleda-Velasquez JF, Quiroz YT, Huentelman MJ, Beach TG, Caselli RJ, Chen Y, Su Y, Myers AJ, Hardy J, Paul Vonsattel J, Younkin SG, Bennett DA, De Jager PL, Larson EB, Crane PK, Keene CD, Kamboh MI, Kofler JK, Duque L, Gilbert JR, Gwirtsman HE, Buxbaum JD, Dickson DW, Frosch MP, Ghetti BF, Lunetta KL, Wang LS, Hyman BT, Kukull WA, Foroud T, Haines JL, Mayeux RP, Pericak-Vance MA, Schneider JA, Trojanowski JQ, Farrer LA, Schellenberg GD, Beecham GW, Montine TJ, Jun GR (2020) Exceptionally low likelihood of Alzheimer's dementia in APOE2 homozygotes from a 5,000-person neuropathological study. Nat Commun 11, 667. [PubMed: 32015339]

[47]. Buccitelli C, Selbach M (2020) mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet 21, 630–644. [PubMed: 32709985]

[48]. Robins C, Liu Y, Fan W, Duong DM, Meigs J, Harerimana NV, Gerasimov ES, Dammer EB, Cutler DJ, Beach TG, Reiman EM, De Jager PL, Bennett DA, Lah JJ, Wingo AP, Levey AI, Seyfried NT, Wingo TS (2021) Genetic control of the human brain proteome. Am J Hum Genet 108, 400–410. [PubMed: 33571421]

[49]. Toikumo S, Xu H, Gelernter J, Kember RL, Kranzler HR (2022) Integrating human brain proteomic data with genome-wide association study findings identifies novel brain proteins in substance use traits. Neuropsychopharmacology 47, 2292–2299. [PubMed: 35941285]

[50]. Yang C, Farias F, Ibanez L, Sadler B, Fernandez MV, Wang F, Bradley J, Eiffert B, Bahena J, Budde J, Li Z, Dube U, Sung YJ, Mihindukulasuriya K, Morris J, Fagan A, Perrin R, Benitez B, Rhinn H, Harari O, Cruchaga C (2020) Genomic and multi-tissue proteomic integration for understanding the biology of disease and other complex traits. medRxiv, doi: 10.1101/2020.06.25.20140277

[51]. Xu J, Mao C, Hou Y, Luo Y, Binder JL, Zhou Y, Bekris LM, Shin J, Hu M, Wang F, Eng C, Oprea TI, Flanagan ME, Pieper AA, Cummings J, Leverenz JB, Cheng F (2022) Interpretable deep learning translation of GWAS and multi-omics findings to identify pathobiology and drug repurposing in Alzheimer's disease. Cell Rep 41, 111717. [PubMed: 36450252]

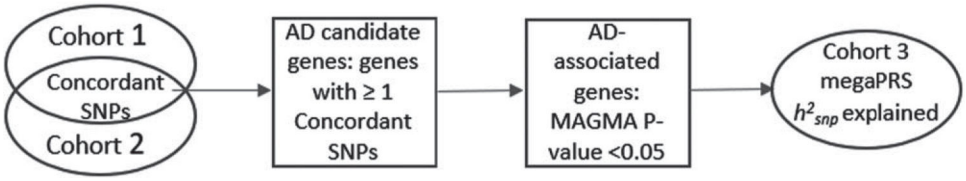

Fig. 1.

Overview of the leave-one-cohort-out (LOCO) strategy. We first used two cohorts to identify concordant SNPs, i.e., SNPs having *p*-values < 0.05 and the same directions of effects in both discovery cohorts. Then we defined AD candidate genes as those having at least one concordant SNP. We further defined AD-associated genes as those AD candidate genes having MAGMA gene-based analysis *p*-values < 0.05. Lastly, we calculated gene-based PRS and evaluated its predictability in the third cohort using megaPRS. We also estimated h^2_{snp} explained in the third cohort.

Table 1

Numbers of genes identified, variation explained by PRS, and h^2_{snp} explained in each LOCO

				***APOE* included**					***APOE* region excluded**				
Discovery dataset	**Target dataset**	**# Genes**	**# SNPs**	R^2	**SE**	h^2_{snp}	**SE**	h^2_{snp} **explained**	R^2	**SE**	h^2_{snp}	**SE**	h^2_{snp} **explained**
UKBB FinnGen	Kunkle	1,745	397,852	0.23	0.03	0.40	0.05	56.12%	0.06	0.02	0.21	0.04	29.93%
Kunkle FinnGen	UKBB	1,711	400,611	0.08	0.06	0.08	0.01	97.46%	0.02	0.05	0.05	0.01	45.79%
Kunkle UKBB	FinnGen	2,003	444,622	0.08	0.04	0.12	0.01	69.53%	0.02	0.04	0.06	0.01	36.06%