Genetically-regulated pathway-polygenic risk score (GRPa-PRS): A risk stratification method to identify genetically regulated pathways in polygenic diseases

Xiaoyang Li [1,2], Brisa S. Fernandes [1], Andi Liu [1,3], Yimei Lu [4], Jingchun Chen [4], Zhongming Zhao [1,3,*], Yulin Dai [1,*]

[1]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[2]Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[3]Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[4]Nevada Institute of Personalized Medicine, University of Nevada Las Vegas, Las Vegas, NV 89154, USA

*To whom correspondence should be addressed:

Yulin Dai, Ph.D.

Center for Precision Health

McWilliams School of Biomedical Informatics

The University of Texas Health Science Center at Houston

7000 Fannin St. Suite E760A Houston, TX 77030

Phone: 713-500-3462

Email: Yulin.Dai@uth.tmc.edu

Zhongming Zhao, Ph.D., FACMI, FAMIA

Center for Precision Health

McWilliams School of Biomedical Informatics

The University of Texas Health Science Center at Houston
7000 Fannin St. Suite E755B Houston, TX 77030
Phone: 713-500-3631
Email: Zhongming.zhao@uth.tmc.edu

ABSTRACT

Background: Alzheimer's disease (AD) is a common neurodegenerative disease in the elderly population, with genetic factors playing an important role. A considerable proportion of elderly people carry a high genetic AD risk but evade AD. On the other hand, some individuals with a low risk for AD eventually develop AD. We hypothesized that unknown counterfactors might be involved in reversing the polygenic risk scores (PRS) prediction, which might provide insights into AD pathogenesis, prevention, and early clinical intervention.

Methods: We built a novel computational framework to identify genetically-regulated pathways (GRPa) using PRS-based stratification for each cohort. We curated two AD cohorts with genotyping data; the discovery and the replication dataset include 2722 and 2492 individuals, respectively. First, we calculated the optimized PRS model based on the three latest AD GWAS summary statistics for each cohort. Then, we sub-grouped the individuals by their PRS and clinical diagnosis into groups such as cognitively normal (CN) with high PRS for AD (resilient group), AD cases with low PRS (susceptible group), and AD/CNs participants with similar PRS backgrounds. Lastly, we imputed the individual genetically-regulated expression (GReX) and identified the differential GRPas between subgroups with gene-set enrichment analysis and gene-set variational analysis in 2 models with and without the effect of *APOE*.

Results: For each subgroup, we conducted the same procedures in both the discovery and replication datasets across three PRS models for comparison. In Model 1 with the *APOE* region, we identified well-known AD-related pathways, including amyloid-beta clearance, tau protein binding, and astrocytes response to oxidative stress. In Model 2 without the *APOE* region, synapse function, microglia function, histidine metabolism, and thiolester hydrolase activity were significant, suggesting that they are pathways independent of the effect of *APOE*. Finally, our

GRPa-PRS method reduces the false discovery rate in detecting differential pathways compared to another variants-based pathway PRS method.

Conclusions: We developed a framework, **GRPa-PRS**, to systematically explore the differential GRPas among individuals stratified by their estimated PRS. The GReX-level comparison among those groups unveiled new insights into the pathways associated with AD risk and resilience. Our framework can be extended to other polygenic complex diseases.

Keywords: genetically-regulated pathways, polygenic risk score, genetically-regulated expression, Alzheimer's disease, resilience, orthogonal effect

INTRODUCTION

Since most common variants of complex diseases only carry small effect sizes, statistical approaches such as polygenic risk scores (PRS) have been developed to combine the effect of multiple genetic variants associated with complex diseases to predict an individual's liability, since a large proportion of complex diseases heritability has been captured by current large-scale GWAS [1]. Accordingly, PRS has become an accurate tool to estimate an individual's genetic risk for a given polygenic disease [1–3], within a certain population. For instance, Alzheimer's disease (AD) is a progressive and debilitating neurodegenerative disease that mainly affects the elderly population. It is characterized by cognitive impairment, memory loss, and behavioral changes, which ultimately lead to a complete loss of independence and death. According to the Alzheimer's Association's 2022 Facts and Figures report, an estimated 6.5 million Americans aged 65 years and older have Alzheimer's dementia, and this number is projected to increase to 13.8 million by 2060 [7]. The prediction performance on pathologically confirmed late-onset Alzheimer's disease (LOAD) cases and controls could reach 0.75–0.84 based on the area under the receiver operating curve (AUROC) [4,5]. Therefore, PRS has the potential to identify individuals at high risk for complex diseases such as AD and potentially guide early intervention, including lifestyle modifications and pharmacological interventions [2,6].

The etiology of AD is complex and multifactorial, involving both genetic and environmental factors. While several risk factors for AD have been identified, genetic factors play a significant role in the development and progression of AD. About 5% of all AD cases develop clinical symptoms before age 65, featured with protein-altering rare mutations on genes, such as amyloid protein precursor (*APP*), presenilin-1 (*PSEN1*) and presenilin-2 (*PSEN2*) genes that cause autosomal dominant forms of early-onset AD [8,9]. For the rest of the 95% late-onset AD (LOAD), large efforts have been conducted to characterize AD associations with common

variants [10–13], rare variants [14,15], and copy number variants [16]. In a previous widely-studied genome-wide association study (GWAS), Kunkle et al. [11] discovered 25 loci of genome-wide significance from 90,000 individuals. Schwartzentruber et al. [13] identified 37 loci by analyzing over 400,000 AD and AD-by-proxy cases and controls. Wightman et al. [12] reported 38 loci by studying a sample of over one million individuals. Among the common variants, the *APOE4* allele is the strongest risk factor, with individuals with this variant showing higher odds for AD with one *APOE4* allele (odds ratio [OR] 4.6 [95% confidence interval (CI) 4.1–5.2]) or two *APOE4* alleles (OR 25.3 [95%CI 20.4–31.2]) compared with those without the *APOE4* allele [17]. However, *APOE* alone does not fully explain the development of LOAD; as a result, ongoing research has been carried out to investigate additional genetic factors and pathways that play a role in both the development and resilience of AD [18–20].

Many elderly people carrying a high AD risk according to their PRS score do not develop AD, while many individuals with a low risk for AD eventually develop AD. This suggests that unknown counterfactors, both genetic and environmental, might not have been fully reflected by the PRS prediction. For instance, individuals with high education attainment had a lower risk of AD and lower cognitive decline, even if they have a high PRS for AD [21]. Several theories [19,21] have been proposed to explain a few resilience metrics that were genetically correlated with these counterfactors and mitigate the effect of genetic risk on AD development. For example, Dumitrescu et al. [19] conducted GWAS for validated metrics of cognitive resilience and identified a genome-wide significant locus near *ATP8B1*. These counterfactors, in the case of genetic factors, are orthogonal factors, and thus unrelated, to the PRS calculation and not merely the opposite of risk [22,23]. In other words, they are truly independent resilience factors. On the other hand, it is still not clear why individuals with AD and low polygenic risk are affected, besides the possible extra rare-variant burden theory [24]. Therefore, these orthogonal genetic factors might provide novel insights into AD pathogenesis and early intervention not only

by identifying biological factors in those who are resilient in high-risk disease group (here called resilient group) but also by identifying underlying biological factors in those who carry an extra burden (not explained by PRS) for the disease.

Recent advances in the study of complex diseases have indicated that integrative analysis of GWAS summary statistics or individual genotype data and various regulatory annotations, including expression quantitative trait loci (eQTL), can provide additional insights not possible with genotyping alone. Methodologies like transcriptome-wide association study (TWAS) can incorporate context-specific expression and show better power to unveil novel disease susceptibility loci [25–29], playing an important role to bridge common variants and their genetically-regulated expression (GReX). Moreover, pathway-level analytical methods can further aggregate the moderate GReX over the genome and increase the power to find interpretable associations [30].

In this work, we designed a novel computational framework, genetically regulated pathways PRS (GRPa-PRS), estimated from imputed GReX, through systematically exploring the differential GReX and GRPa among individuals by their disease risk strata in two AD independent cohorts. Specifically, we aimed to 1) explore GRPa related to different AD risk strata; 2) disentangle the GRPa among different AD risk strata that is independent of the dominant factor of *APOE*; 3) identify the GRPa that is orthogonal to their AD risks, which potential unveils resilience and extra burden, or susceptibility, factors related to AD; 4) benchmark with other pathway methods to evaluate the performance of our framework. Moreover, our GRPa-PRS framework can be further expanded to other complex polygenic traits to explore their GRPas.

MATERIALS AND METHODS

GWAS datasets

Several different AD GWAS summary statistics were used to calculate the PRS. We utilized three recent AD and AD-related case-control GWAS summary statistics that contained both shared and distinct samples to investigate and assess potential variations in strata groups derived from GWAS weight during PRS calculation. The first dataset, Kunkle's (K) meta-analysis of AD GWAS, focuses on the phenotype LOAD with the age of onset > 65 years, which includes 11,480,632 variants from the study consisting of 21,982 AD cases and 41,944 CN controls [11]. The second GWAS, Schwartzentruber's (S) meta-analysis of GWAS for AD and AD by proxy phenotypes (GWAX) in the UK BioBank (UKBB), is composed of 898 AD cases, 52,791 AD by proxy cases, and 355,900 controls including a total of 10,687,077 SNPs [13]. The last dataset is from Wightman (W) et al. [12] meta-analysis of AD GWAS. We adapted the summary statistics of their GWAS meta-analysis without the proxy cases from the UKBB and 23andMe, which resulted in 39,918 cases and 358,140 controls including a total of 12,674,019 SNPs [12].

Genotyping data composition in discovery and replication cohorts

We collected our discovery and replication cohorts from (https://www.ncbi.nlm.nih.gov/gap/) and synapse (https://www.synapse.org/), respectively. The discovery cohort is comprised by the National Institute of Aging/ Late Onset Alzheimer's Disease Study (NIA/LOAD) cohort consents 1 and 2 (ADc12) [phs000168] [31] and the Multi-Site Collaborative Study for Genotype-Phenotype Associations in Alzheimer's Disease (GenADA) [phs000192] [32]. Raw SNP arrays were downloaded from dbGaP accordingly (accessed on 6/15/2021). The replication cohort includes the Whole Genome Sequencing (WGS) (https://www.synapse.org/#!

Synapse:syn5550382) data from the Religious Orders Study and Memory and Aging Project (ROS/MAP) Study [33], the MayoRNAseq (Mayo) study [34], and the Mount Sinai Brain Bank (MSBB) study [35] and imputed SNP array (https://www.synapse.org/#!Synapse:syn3157325) data from ROS/MAP. The raw SNP array from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [36] was downloaded from https://adni.loni.usc.edu (accessed on 6/15/2021).

Imputation of genotyping dataset

To boost the SNPs that are not directly genotyped in the raw genotyping array, we adapted the previous pipeline to impute these raw AD genotyping data [37]. Briefly, we performed the standard variants checking procedure using the tools developed by McCarthy group [38]. Imputation was conducted for ADc12, GenADA, and ADNI datasets using the Michigan Imputation Server (minimac4) [39] with the 1,000 Genome Phase 3v5 European Panel, and SNP with $r^2 < 0.6$ was filtered. The imputed genotyping data was downloaded from Synapse, consisting of individuals from the CHOP batch (382) in ROS/MAP study. From the method description (SynID: syn3157325), the SNPs which passed quality control as described in De Jager et. al [40] were imputed using BEAGLE software, version 3.3.2, with the reference haplotype panels from 87 Centre d' Etude du Polymorphisme Humain individuals (CEPH) of Northern European ancestry in the 1000 Genomes Project (1000 Genomes Project Consortium interim phase I haplotypes, 2010-2011 data freeze).

Diagnosis criteria

In this study, we defined AD cases based on the clinical diagnosis labels in each cohort. Specifically, for the discovery cohort, individuals with dementia who were diagnosed with definitive, probable, or possible Alzheimer's disease at any stage of their clinical course, according to the criteria introduced in 1984 by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders

Association (NINCDS-ADRDA), were categorized as AD cases [41]. Controls were cognitively normal (CN) individuals. Those with unspecified dementia, unconfirmed controls, or other neurological diseases were excluded.

For the replication cohort, we assigned the clinical diagnosis labels to each participant in different sub-cohorts depending on the following criteria. The clinical diagnosis labels were assigned based on cognitive function evaluation and factors available in each cohort. As for the discovery dataset, participants in the ROS/MAP cohort with a consensus diagnosis of AD with no other cause of cognitive function impairment were classified as AD cases, while individuals with no cognitive function impairment were classified as CN controls [42]. Diagnosis in the last follow-up visit was used for clinical diagnosis labeling if a final consensus diagnosis was not available. Individuals in the MSBB cohort were classified based on Clinical Dementia Rating (CDR, range 0 to 5). Participants with CDR ≤ 0.5 were classified as controls, and those with CDR ≥ 3 were classified as AD cases [35]. Participants in the Mayo cohort were classified based on existing clinical diagnosis labels. However, those with a clinical diagnosis of pathological aging or progressive supranuclear palsy diagnosis were removed. For the ADNI cohort, AD cases and controls diagnosis were based on the last recorded diagnosis status in the metadata.

Genotyping data integration and quality control for GWAS

We followed the diagnosis criteria described in the previous section to label the participants and extracted the clinically diagnosed AD and CN individuals for both discovery and replication cohorts. We combined the imputed genotyping datasets (ADc12 and GenADA) with VCFtools [43] and formed the discovery cohort (ADc12ADA). We removed non-European ancestry individuals according to their PCA relative to 1000 Genome individuals [44]. We annotated rsid and European allele frequency for both the discovery and validation dataset using ANNOVAR

[45] and annotated the rsid column of VCF file using BCFtools [46]. KING2.2 [47] was used to detect relativeness between individuals. We removed individuals with relatives closer than second degree within each dataset. Then, we conducted the following variant-based quality control using VCFtools [43], BCFtools [46], and PLINK [48,49], including removed insertion, deletions, and multi-allelic variants; excluded variants without rsid; excluded variants with European allele frequency < 0.01; filtered variants whose call rate were less than 98%; filtered individuals with call rate less than 95%. In addition, we excluded the *APOE* region (chr19, 44.4M-46.5M) to create a filtered genotype version for Model 2 of GRPa-PRS framework, which will be used to impute gene expression without the *APOE* region gene. After the quality control process, the final discovery cohort (ADc12ADA) included 2722 individuals (1337 cases and 1385 controls) with 7,232,347 SNPs. Their demographic information and *APOE* genotype of participants are listed in **Table 2**.

Similarly, we used VCFtools [43] to merge the synapse whole genome sequencing (WGS) [ROS/MAP (1196), Mayo (346), MSBB(345)], imputed Synapse genotyping dataset [CHOP (382)] and ADNI dataset (1065) and formed the replication cohort (SynapseADNI). Among the above participants, only cases and controls were included. We used the same procedure in the discovery cohort to conduct the quality control for the genotyping data. The final replication cohort (SynapseADNI) included 2492 individuals (1338 cases/1154 controls) with 6,501,459 SNPs. Their demographic information and *APOE* genotype are listed in **Table 2**.

PRS and subgroups

We used the overlapping variants between HapMap3 Project [50] and GWAS summary statistics to match the variants in genotyping data of both cohorts. LDpred2 [51] (-grid model) was applied to calculate candidate PRS for each individual with each hyperparameter combination using matched variants. The hyperparameters include heritability (h^2), probability of

causal variant (p), and sparse (binary variable which shrinks some coefficients to 0). Specifically, we first estimated the h^2 by LD score regression by `bigsnpr` R package [52]. Then, the combination of multiple hyperparameters was calculated using the estimated h^2 (0.3 * h^2, 0.7 * h^2, 1 * h^2, 1.4 * h^2), probability of causal variant p ranging from 0.001 to 1, as well as the binary variable sparse (TRUE or FALSE). By using the clinical diagnosis label (Diagnosis criteria section) as the ground truth, the hyperparameter combination that reaches the highest area under the curve (AUC) was selected to generate the best PRS model for the according cohort. We generated the best PRS model for different variant p-value thresholds (1, 0.5, 0.2, 0.1, 0.05). For example, when the variant p-value threshold was set at 0.5, only variants with p-values less than 0.5 from the summary statistics were included in the PRS calculation. Their prediction performance was compared using the AUC metrics. The PRS reaching the highest AUC was used to subgroup individuals. The detailed subgrouping strategy is shown in **Table 1** and **Fig. 2A**. Except for the clinical diagnosis group (Diag) comparison between cases and controls, we repeated the above procedure for three different GWAS summary statistics (Kunkle et al.(K) [11], Schwartzentruber et al.(S) [13], and Wightman et al.(W)) [12] mentioned in the GWAS datasets section and generated 3 sets with different labels. Notably, to ensure meaningful downstream statistical analyses, 15 quantiles were set for comparison groups in discovery and replication cohorts to reach the least sample size 50 in each comparison group (**Table S1-S6**).

Gene set curation

We obtained the gene set from four different resources, including AD-related pathways and terms and brain cell-type-specific pathway (AD brain function), non-redundant gene ontology (GO), and canonical pathways. We limited the gene number in the gene set ranging from 5 to 500. Specifically, AD brain function gene sets contain 167 gene sets, curated from two major resources: 70 AD-related function curation [53–55] and 97 brain cell-type-specific functions

(https://ctg.cncr.nl/software/genesets (accessed on 6/9/2020) [56–58]). GO curation (Biological Process: BP, Molecular Function: MF, and Cellular Component: CC) is composed of 1303 non-redundant GO terms from WebGestalt [59] (accessed on 5/7/2021). Lastly, we curated 2082 canonical pathways from three major resources, Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, and BioCarta pathways from the Molecular Signatures Database [60] (MSigDB 7.4 C2 category, accessed on 5/10/2021). Overall, we curated 4202 unique functional gene sets.

Inferring the GReX for individuals and association study with AD

We used PrediXcan [61] to calculate the genetically regulated expression (GReX) via the Multivariate Adaptive Shrinkage in R (MASHR) model for all the individuals and their genotyping data after QC (**Table 2**). The imputed gene expression was calculated for 13 different brain regions. These served as the input for two differential GRPa approaches: 1) **GRPa-GSEA**: transcriptome-wide-association study (TWAS) + gene-set enrichment analysis (GSEA) and 2) **GRPa-GSVA**: gene set variational analysis (GSVA) +logistic likelihood ratio test (LogisticLRT).

Genetically-regulated pathways gene set enrichment analysis (GRPa-GSEA)

For the GRPa-GSEA approach, MultiXcan [62] was used to integrate the measurements across brain regions and identify the significant genes via the F-test to assess the joint fit significance of the regression coefficients, where the NULL model only includes demographic features as covariates; The covariates include age (age_proxy), sex, and the top 5 principal components of their genotype; Model 1 includes both demographic features and predictor derived from full genotype as predictors; Outcome is defined for each stratum with PRS and clinical diagnosis of individuals as shown in **Table 1**; GReX per gene (full) is obtained from PrediXcan with all the genes. For each gene, gene expression in different tissues was integrated by PCA and treated as predictors in the model.

NULL Model: Outcome ~ covariates. (1)

Model 1: Outcome ~ GReX per gene (full) + covariates. (2)

Additionally, due to the dominant effect of *APOE*, we designed Model 2 (the no-*APOE* model), which includes both demographic features and predictors derived from the genotype excluding *APOE* region as the predictor. The GReX of Model 2 was generated by utilizing genotype data that excluded the *APOE* region and estimated the joint fit significance of regression coefficients, which is suggested to be independent of *APOE* effect. Similarly, the outcome is defined for each stratum with PRS and clinical diagnosis of individuals as shown in **Table 1**. The same covariates were applied in Model 2 as in Model 1.

Model 2: Outcome ~ no-*APOE*-GreX per gene (no *APOE* region gene) + covariates. (3)

Based on the results of MultiXcan, we obtained the list of genes with their p-values for the joint fit and calculated the $-\log_{10}$(p-value) to generate the significant gene list. WebGestaltR GSEA was conducted using the rank of the full gene list with their $-\log_{10}$(p-value) to identify significant GPRa for each subgroup [59]. We conducted the GSEA for the three aforementioned gene sets/pathways resources, AD brain function, GO, and canonical pathway. We defined the significant GRPa adjusted $p < 0.05$ after Benjamini-Hochberg (BH) test for each gene set curation.

Genetically-regulated pathways gene set variation analysis (GRPa-GSVA)

Gene set enrichment was performed for all the imputed AD individuals using the R package 'GSVA' (function gsva - arguments: method = "gsva", mx.diff = TRUE) [63]. GSVA implements a non-parametric unsupervised method of gene set enrichment that allows an assessment of the relative enrichment of a selected pathway across the individual space. The scores of each GSVA pathway from all individuals followed an approximately normal distribution, which

depicted the magnitude difference between the largest positive and negative random walk of each term or pathway. We calculated the aforementioned gene sets (with minimum 5 genes and maximum 500 genes) by using the GReX for each individual and each of 13 brain regions. We proposed two models as GRPa-GSEA models to detect the conditions with or without *APOE* region GReX. We used logisticLRT to assess the increment in goodness of fit (deviance of NULL model – deviance of Model 1) and (deviance of NULL model - deviance of Model 2) for each genetically regulated pathway in each tissue by the full GRPa (Model 1) and no-*APOE*-GRPa Model2, where outcome and covariates were the same as mentioned in the previous model, and GRPa was calculated from GSVA for each tissue and each individual from GReX input.

Model 1: Outcome ~ GRPa (full) + covariates. (4)

We defined the significant GRPa after Bonferroni correction for each gene set curation.

For Model 2, the same procedure and covariates were conducted for the model using GReX without the *APOE* region gene as follows:

Model 2: Outcome ~ no-*APOE*-GRPa (no *APOE* region gene) + covariates. (5)

PRSet pathway polygenic risk score test

To evaluate the performance of our GRPa-PRS, we benchmarked the pathway-level PRS estimated by one of the latest methods, PRSet [64]. Briefly, PRSet employs the classical approach of clumping and thresholding (C+T) to compute personalized PRSs for specific genomic pathways and terms of interest using individual genotyping information. SNPs from a 250kb window for up- and down- stream of each gene coordinate from the curated gene set were accumulated. To evaluate the association between each pathway and AD outcome, a logistic regression model was applied, which considered covariates such as sex, age

(age_proxy), and the first 5 genotype principal components; the same number of random SNPs were permuted 10,000 times as the background to assess the empirical competitive *p*-values. The BH test was used to correct for multiple testing in the gene sets. We used the number of pathways passing the false discovery rate (FDR) < 0.05 threshold to estimate the performance of GRPa-PRS and PRSet for all three curated gene sets between the two cohorts. Similarly, as we did in the two approaches GRPa-GSEA and GRPa-GSVA, we conducted logistic regression for both genotyping data with/without SNPs within the *APOE* region to evaluate PRSet's performance in identifying signals independent of the *APOE* region. The same outcome and covariates were applied as the previous models.

Model 1: Outcome ~ gene set PRS (full) + covariates. (6)

Model 2: Outcome ~ gene set PRS (no *APOE* region SNPs) + covariates (7)

Comparing the variance explained by GRPa-GSVA vs. PRSet

We introduce GRPa generated by our GRPa-GSVA into predictors of formula (6) to compare the improvement that comes with PRSet and GRPa generated by our GRPa-GSVA method as shown in formula (8). The same outcome and covariates were applied as the previous models. To evaluate the goodness of fit in the logistic regression model, we adopt Nagelkerke's R squared to compare the variance explained by the formula (1), formula (6) and formula (8) below, where the Nagelkerke's R squared of formula (1) (NULL model) is the denominator and the Nagelkerke's R squared of formula (6) and formula (8) are nominators, respectively. We obtain the R2 improvement ratio among different enrichment databases.

Outcome ~ gene set PRS (full) + GRPa (full)+ covariates. (8)

Orthogonality test for significant GRPa among resilient and susceptible groups

To disentangle the GRPa that potentially contains resilience and susceptibility effects, we conducted a Pearson correlation test to assess whether the GRPas are significantly correlated with the PRS. Specifically, we defined the orthogonal effect as the differential GRPa that does not have a significant correlation with the PRS score. Subsequently, the resilience effect is the significantly differential orthogonal GRPa identified in the TB15Ctr and T15 groups (Table 1), while the susceptibility effect is the significantly differential orthogonal GRPa identified in TB15AD and B15 groups (**Table 1**).

RESULTS

Framework of GRPa-PRS and PRS AD risk strata strategy

Individuals carry different genetic liabilities to certain traits and diseases. However, which and how genetic risks play their roles in the pathway and individual level has not been fully revealed. The PRS-based methods have been widely used to characterize such liabilities. We designed a computational framework (**Fig. 1**), genetically-regulated-pathway polygenic risk score (GRPa-PRS), that can assess the individual-level pathway and terms risk from genetically-regulated-expression (GReX) and further compare the GRPas among biologically meaningful PRS-estimated risk strata (**Fig. 2A**). We designed two statistical approaches to identify the differential GRPas among those strata. Lastly, we benchmark our model performance with another recently published method, PRSet [64].

We used AD as an exemplar of our framework as AD is a polygenic disease. We curated two independent AD cohorts (ADc12ADA with 2722 individuals and SynapseADNI with 2492 individuals). We adapted the clinical diagnosis as the case and control labels. After, we

estimated the AD PRS risk for each individual using the effect size from the previous three large-scale AD GWAS summary statistics [11–13]. Their PRS model was optimized using LDPred2 for three different GWAS summary statistics parallelly to ensure a less biased estimation. The PRS AD risk strata were defined in **Table 1** and illustrated in **Fig. 2A.** Intuitively, for the case_control (all cases vs. all controls) and TB15all (all top 15% and all bottom 15%) groups, we explored the GRPas difference among the individuals from AD diagnosis groups in full cases or controls or extreme cases or controls. These GRPa differences mainly represent the PRS-burden difference among comparison groups when considering their diagnosis label. In the TB15AD (top15% cases vs bottom15% cases) and TB15Ctr (top15% controls vs bottom15% controls), we explored the high-risk burden group vs low-risk burden group in AD cases and controls. These GRPa differences could be mainly explained by the PRS-burden within the same diagnosis label. Lastly, in T15 group (AD cases vs controls in high PRS group) and B15 group (AD cases vs Controls in low PRS group), we explored the GRPa differences in individuals with similar PRS burden but different diagnoses, suggesting the resilience and extra burden related GRPas.

Table 1. Subgroup strategy and comparison group interpretation in the discovery and replication data sets

Subgroup abbreviation	Comparison group 1 in cohort (discovery/replication)	Comparison group 2 in cohort (discovery/replication)	Interpretation for GRPa difference in each comparison
case_control group Fig. 2A (1)	All clinically diagnosed cases	All clinically diagnosed controls	General GRPa difference between cases and controls
TB15all group Fig. 2A (2)	Individuals with the top 15% PRS among total participants *	Individuals with the bottom 15% PRS among total participants *	PRS burden difference explained by GRPa in extreme quantile between cases and controls
TB15AD	Individuals with the top	Individuals with the	Extra burden effect and

group Fig. 2A (3)	15% PRS among total cases **	bottom 15% PRS among total cases ** (**susceptible group**)	PRS burden explained by GRPa difference among high-risk cases and low risk cases
TB15Ctr group Fig. 2A (4)	Individuals with the top 15% PRS among total controls *** (**resilient group**)	Individuals with the bottom 15% PRS among total controls ***	Resilience effect and PRS burden explained by GRPa difference among high-risk controls and low-risk controls
T15 group Fig. 2A (5)	Cases within the top 15% PRS among all participants starting from the top 2.5 quantiles of outlier controls	Controls within the top 15% PRS among all participants starting from the top 2.5 quantiles of outlier controls (**resilient group**)	Resilience effect (mainly) explained by GRPa difference among similar high PRS cases and controls
B15 group Fig. 2A (6)	Cases within the bottom 15% PRS among all participants excluding the bottom 2.5 quantiles of outlier cases (**susceptible group**)	Controls within the bottom 15% PRS among all participants excluding the bottom 2.5 quantiles of outlier cases	Extra burden effect (mainly) explained by GRPa difference among similar low PRS cases and controls

Abbreviations: AD, Alzheimer's disease; B, bottom; T, top; TB, top vs bottom; case_control, clinical diagnosis. *: Excluding outliers of all with PRS lie outside of the range from 2.5 to 97.5 quantile of all participants, **: Excluding outliers of cases with PRS lie outside of the range from 2.5 to 97.5 quantile of cases, ***: Excluding outliers of controls with PRS lie outside of range from 2.5 to 97.5 quantile of controls.

PRS model performance evaluation and selection

We calculated the PRS based on three different GWAS summary statistics (S/K/W) for both discovery and replication cohorts. Using the clinical diagnosis as the ground truth, we determined the tuning parameters (h^2, p, sparse) by their prediction performance (**Fig. 2B and Fig. 2C**). The same PRS calculation process was performed for a list of variant p-value thresholds (1, 0.5, 0.2, 0.1, 0.05). Prediction performance of the PRS including the *APOE* region (Model 1) had the highest AUC with variant p-value threshold at 0.5 for both the discovery

dataset (AUC (S/K/W): 0.90/0.95/0.92) and replication dataset (AUC(S/K/W): 0.81/0.83/0.81). For Model 2, the PRS with no-*APOE* region, the highest AUC was reached when setting the variant p-value threshold at 0.5. The exclusion of the *APOE* region in the PRS calculation led to a slight decrease in the AUC (discovery AUC(S/K/W): 0.88/0.94/0.90; replication AUC(S/K/W): 0.79/0.81/0.78). Therefore, PRSs calculated under 0.5 were used to generate the subgroup indicator by the strategy mentioned in Methods, their generated PRSs distribution is shown in **Fig. S1**. The demographic information for each subgroup is shown in **Table S1-6.**

***APOE*-related pathways dominate the differential GRPas in diagnosis groups and extreme diagnosis groups.**

As summarized in **Table 1**, we first explored the differential GRPas in PRS burden in the case_control group (all cases and controls) and TB15all group (extreme quantile cases and controls). In the GRPa-GSEA Model 1 result (**Fig. 3A**), the discovery cohort revealed 20 significant GO pathways and terms when comparing cases and controls. The top three terms were CC 'protein lipid complex' (FDR: 1.15×10^{-2}), BP 'synaptic transmission cholinergic' (FDR: 1.21×10^{-2}), and MF 'proteoglycan binding' (FDR: 1.35×10^{-2}). Within the three TB15all subgroups, 32 significant GO pathways were identified. The top three pathways were CC 'protein lipid complex' (FDR: 1.10×10^{-3}), BP 'protein lipid complex subunit organization' (FDR: 5.00×10^{-3}), and BP 'cellular response to lipoprotein particle stimulus' (FDR: 7.84×10^{-3}). The replication cohort (**Fig. 3B**) replicated ~50%-78.6% of the Model 1 differential GRPas from the discovery cohort, as shown in the Venn plot (**Fig. 3E**). However, for our no-*APOE* Model 2 (**Fig. 3C** and **Fig. 3D**), no significant pathway was identified in the discovery cohort, suggesting that *APOE* and its related functions dominate the PRS-related GRPas. In the replication cohort, only two terms were identified: CC 'mast cell granule' (FDR: 1.06×10^{-2}) and MF 'pattern recognition receptor activity' (FDR: 3.45×10^{-2}).

In the GRPa-GSVA GO analyses, the discovery cohort revealed 21 significant GO pathways when comparing cases and controls (**Fig. 4A**). The top three pathways and terms were BP 'amyloid beta clearance' (p-value: 1.08×10^{-30}), MF 'lipoprotein particle receptor binding' (p-value: 1.51×10^{-25}), and BP 'protein containing complex remodeling' (p-value: 2.60×10^{-17}). Within the three TB15all subgroups, 26 significant GO pathways were identified. The top three pathways and terms remained the same: BP 'amyloid beta clearance' (p-value: 4.20×10^{-30}), MF 'lipoprotein particle receptor binding' (p-value: 4.26×10^{-24}), and BP 'protein containing complex remodeling' (p-value: 4.65×10^{-16}). The Venn plot (**Fig. 4E**) demonstrated partial replication of Model 1 results from the discovery cohort in the replication cohort (**Fig. 4B**). Consistent with the GRPA-GSEA findings, Model 2 revealed a limited number of pathways, as shown in **Fig. 4C** and **Fig. 4D**. In the discovery cohort, within the TB15all subgroup comparison, BP 'tube formation' (p-value: 3.33×10^{-6}) and MF 'sulfur compound binding' (p-value: 1.36×10^{-5}) were identified. Conversely, in the replication cohort, the BP 'positive regulation of neurogenesis' (p-value: 8.70×10^{-6}) emerged. These pathways identified in Model 2 may be independent of the *APOE* effect.

We expanded our analysis to include two additional gene set curations (AD brain function terms and canonical pathways, see Methods) by GRPa-GSEA (**Fig. S2** and **Fig. S3**), as well as GRPa-GSVA (**Fig. S4** and **Fig. S5**). The results exhibited a similar pattern to the GO pathways analysis, wherein a limited number of pathways were identified in the no-*APOE* model compared to the comprehensive model. This finding suggests that the impact of *APOE* is prominent. Notably, in the no-*APOE* Model 2, significant pathways and terms related to microglia and synapses were detected, including ‘microglia antigen processing’, ‘microglia (NF) kappa B kinase’, GO CC 'synapse', and REACTOME 'phenylalanine and tyrosine metabolism'. These pathways and terms are suggested to play their roles independently of the *APOE* directly involved functions.

TB15AD group has more PRS-related GRPas than TB15Ctr does despite their similar PRS burden difference.

Then, we explored the GRPa difference in PRS burden difference in TB15AD group (high-risk cases and low-risk cases) and TB15Ctr group (high-risk controls and low-risk controls). In the GRPa-GSEA analysis of the discovery cohort (**Fig. 3A**), we observed 13 significantly different GRPas in the comparisons within the TB15AD group, while only 2 pathways and terms were found in the TB15Ctr group. Similar trends were observed in the results of the other two curations (**Fig. S2** and **Fig. S3**). These suggested that TB15AD contained more PRS-related GRPas than TB15Ctr groups despite the comparison groups sharing the same PRS burden difference. The top three significant GO pathways and terms identified in the TB15AD group were BP 'synaptic transmission cholinergic' (FDR: 6.50×10^{-3}), MF 'lipoprotein particle receptor binding' (FDR: 7.00×10^{-3}), and BP 'response to lipoprotein particle' (FDR: 8.35×10^{-3}). Notably, the replication cohort (**Fig. 3B**) successfully replicated model 1 results of TB15AD from the discovery cohort, as shown in **Fig. 3E**. For TB15Ctr, there are only two pathways identified in the discovery dataset, namely 'exploration behavior' (FDR: 1.70×10^{-3}) and 'behavioral defense response' (FDR: 2.22×10^{-2}), which is related to specific behavior in response to a novel environment or stimulus that involves nervous system activity [65]. In our no-*APOE* model (**Fig. 3C** and **Fig. 3D**), no significant GO term was identified. However, when exploring AD brain function gene set curation (**Fig. S1D**), we detected significant GO terms CC 'neurofibrillary tangle' and GO BP 'myelin maintenance' in the TB15AD groups, which survives multiple test corrections. Furthermore, in the analysis of canonical pathways (**Fig. S2D**), we identified significant BIOCARTA 'AHSP pathway' and Reactome 'activated NTRK2 signals through RAS' pathways in the TB15AD groups.

In the GRPa-GSVA analysis of the discovery cohort (**Fig. 4A**), we found 3 significantly different GRPas and terms in comparisons within the TB15AD group: MF 'lipoprotein particle receptor

binding' (p-value = 1.14×10^{-8}), BP 'amyloid beta clearance' (p-value: 7.82×10^{-8}), and BP 'response to tumor cell' (p-value: 1.59×10^{-5}). In the TB15Ctr group, we identified 4 pathways and terms: MF 'thiolester hydrolase activity' (p-value: 1.50×10^{-6}), BP 'response to tumor cell' (p-value: 5.33×10^{-6}), BP 'amyloid beta clearance' (p-value: 1.54×10^{-5}), and MF 'carbohydrate transmembrane transporter activity' (p-value: 3.25×10^{-5}). In the replication cohort (**Fig. 4B**), only the MF 'lipoprotein particle receptor binding' pathway in the TB15AD group was successfully replicated. In the no-*APOE* model, the two terms were identified only in the TB15Ctr groups of the discovery dataset: MF 'thiolester hydrolase activity' (p-value: 1.54×10^{-6}) and MF 'carbohydrate transmembrane transporter activity' (p-value: 3.24×10^{-5}). The number of signals detected from the other two curations (**Fig. S4** and **Fig. S5**) are mainly enriched in Synapase, supporting our previous findings.

T and B groups comparison highlights a few potential resilience and extra burden-related GRPas that is different in PRS-controlled groups.

Lastly, we assessed the potential differential GRPas in PRS burden balanced groups: T15 group (high-risk cases and low-risk cases) and B15 group (high-risk controls and low-risk controls), suggesting the resilience effect and extra burden effect, respectively. In the GRPa-GSEA analysis (**Fig. 3A**), no significant differential GRPas were found in the discovery cohort for the T15 and B15 groups. In the replication cohort, we identified one significant term, CC 'dense core granule' (FDR: 4.78×10^{-2}). This suggests that despite similar PRS burdens, the GRPas may be differently associated with AD outcomes. In the no-APOE model (**Fig. 3C and Fig. 3D**), no significant GO signals were identified. However, in the other two gene set curations (Supplementary **Fig. S2-S5**), we discovered GOCC 'gamma secretase complex' (**Fig. S2D**) and Reactome 'formation of apoptosome' **(Fig. S3D**) in the T15 group, which might play important role in AD resilience and independent of the *APOE* effect.

Interestingly, in the GRPa-GSVA GO analysis (**Fig. 4A**), the MF 'thiolester hydrolase activity' was detected in both model 1 (p-value: 1.11×10^{-5}) and model 2 (p-value: 1.12×10^{-5}) for the KT15 group based on the discovery cohort. Notably, this term is also highlighted in the previous STB15Ctr group, which suggests this differential GRPa mainly comes from the same high-risk control group (resilience group) even from different PRS measurements. No signals were identified in the T15 or B15 groups in the replication cohort. In the other two gene set curations, we found GOCC 'inhibitory synapse' in the B15 group (**Fig. S4C**) and REACTOME 'histidine catabolism' in the T15 group (**Fig. S5C**), which might be independent of the APOE effect.

Shared and unique significant pathways and terms that are identified by three approaches.

We assessed the performance of our GRPa-PRS by comparing it to PRSet, a method that computes pathway PRSs based on SNPs and personalized gene sets for each individual. To evaluate pathway-level significance, we calculated FDR-adjusted empirical p-values in PRSet using subgroup indicators as phenotypes and adjusting for the same covariates. We compared the union set of pathways identified using three GWAS summary statistics in the discovery and replication datasets across different methods. As shown In **Fig. 5A**, we observed that the number of overlapping pathways identified by GRPa-GSEA and GRPa-GSVA was comparable, while PRSet identified a much larger number of pathways and terms. For example, out of 1303 non-redundant GO terms, PRSet identified 175 pathways and terms in the discovery dataset and 154 pathways and terms in the replication dataset when using full genotype information (see **Fig. S6**). PRSet identified more than 10% of the curated gene sets, which suggests a possible issue of inflated type 1 error rates. The other 2 databases of PRSet showed a similar trend (**Fig. S7-S8**). Therefore, we focused on the results obtained from our GRPa-GSEA and GRPa-GSVA (**Fig. 5B**)**.** We found 19 overlapping signals between these two methods in the case vs. control comparison and 21 overlapping signals in the TB15all groups. In the TB15AD

group, only 3 signals overlapped, while there were no overlapping signals in the remaining groups. These findings underscore the comparable performance within our two methods.

Comparing the variance explained by GRPa-GSVA vs. PRSet

As described in the methods section, we assessed the improvement brought by PRSet and GRPa-GSVA by comparing Nagelkerke's *R* square values of formula 6 and formula 8 with that of formula 1. Based on the discovery dataset, the results are presented in **Fig. 5C**. Taking the GO enrichment database as an example, in formula 6, which includes pathway-specific PRS as a covariate, the ratio of *R* square values ranged from 1 to 1.98 (mean: 1.14). In formula 8, where both pathway-specific PRS and GRPa generated by GSVA were included as covariates, the ratio of *R* square values ranged from 1.00 to 2.01 (mean: 1.17). These findings suggest that formula 8 exhibits a significantly better performance (t-test p-value: 1.68×10^{-3}) compared to formula 6. Similar patterns were observed in the other two gene set curations in **Fig. 5C**.

Definition of orthogonal effect in the resilient and susceptible groups

In this work, we aimed to identify a novel GRPa that is independent of PRS risk and has resilience or susceptibility effects to AD, as we defined in the orthogonal effect GRPas session (Methods). Therefore, these resilience and susceptibility GRPas should have the following features: 1) to be significant in the TBCtr or T15 groups for the resilience effect, and in the TBAD or B15 groups for the susceptibility effect (**Table 1**); 2) to be orthogonal instead of opposite to the PRS (not significantly correlated with PRS risk). In **Fig. 6**, we illustrated a few pathways and terms as positive and negative examples. For instance, MF 'thiolester hydrolase activity' was a notable signal identified in the TB15Ctr group and the T15 groups. More importantly, it is significant in Model 2 as well. As shown in **Fig. 6A,** there is no significant correlation between its term score and PRS within the brain frontal cortex BA9, suggesting the MF 'thiolester hydrolase activity' has an orthogonal effect to PRS effect and is a resilience

GRPa given it is significant in Model 2 (**Fig. 4C**). Similarly, for MF 'carbohydrate transmembrane transporter', another prominent signal identified in the TB15Ctr group, there is no significant correlation with PRS within the brain caudate basal ganglia, as shown in **Fig. 6C**. On the contrary, BP 'amyloid beta clearance', a representative signal identified in the case_control group, TB15all groups, TB15AD groups, and TB15Ctr groups, demonstrated a significant correlation ($r = 0.13$; $p = 1.058 \times 10^{-11}$) with the PRS. As depicted in **Fig. 6E**, BP 'amyloid beta clearance' is positively correlated with PRS score and significantly different among the comparison groups as shown in the heatmap in **Fig. 4A**. The MF 'thiolester hydrolase activity' only exhibited a significant difference ($p < 0.1$, two-sample t-test) in the STB15Ctr group and KT15 group (**Fig. 6B**), which replicates our results from **Fig. 4A**. Similarly, MF 'carbohydrate transmembrane transporter' showed a significant difference ($p < 0.1$, two-sample t-test) only in the WTB15Ctr group (**Fig. 6D**), further supporting our earlier findings. These two terms demonstrate distinct patterns compared to BP 'amyloid beta clearance', which displayed a significant difference in the STB15all group, STB15Ctr group, and WTB15Ctr group, but no significant signals in the KT15 group (**Fig. 6F, Fig. 4A**).

Table 2 Demographic summary statistics for the Discovery dataset and Replication dataset

	Discovery			Replication			
	Cases (N=1337)	**Controls (N=1385)**	**Total (N=2722)**	**Cases (N=1338)**	**Controls (N=1154)**	**Total (N=2492)**	**Overall (N=5214)**
Age							
Mean (SD)	81.0 (7.72)	74.9 (7.59)	77.9 (8.24)	81.7 (8.41)	81.2 (7.71)	81.5 (8.09)	79.6 (8.36)
Median [Min, Max]	81.4 [57.3, 103]	75.0 [57.0, 98.0]	78.0 [57.0, 103]	83.8 [55.0, 95.0]	82.2 [57.0, 96.0]	83.0 [55.0, 96.0]	80.0 [55.0, 103]
Sex							
Male	531 (39.7%)	550 (39.7%)	1081 (39.7%)	596 (44.5%)	466 (40.4%)	1062 (42.6%)	2143 (41.1%)
Female	806 (60.3%)	835 (60.3%)	1641 (60.3%)	742 (55.5%)	688 (59.6%)	1430 (57.4%)	3071 (58.9%)
APOE SNP Genotype							

	Discovery			Replication			
	Cases (N=1337)	**Controls (N=1385)**	**Total (N=2722)**	**Cases (N=1338)**	**Controls (N=1154)**	**Total (N=2492)**	**Overall (N=5214)**
APOE2/APOE2	1 (0.1%)	6 (0.4%)	7 (0.3%)	3 (0.2%)	7 (0.6%)	10 (0.4%)	17 (0.3%)
APOE2/APOE3	42 (3.1%)	170 (12.3%)	212 (7.8%)	72 (5.4%)	160 (13.9%)	232 (9.3%)	444 (8.5%)
APOE2/APOE4	40 (3.0%)	28 (2.0%)	68 (2.5%)	26 (1.9%)	17 (1.5%)	43 (1.7%)	111 (2.1%)
APOE3/APOE3	412 (30.8%)	896 (64.7%)	1308 (48.1%)	601 (44.9%)	733 (63.5%)	1334 (53.5%)	2642 (50.7%)
APOE3/APOE4	658 (49.2%)	265 (19.1%)	923 (33.9%)	522 (39.0%)	223 (19.3%)	745 (29.9%)	1668 (32.0%)
APOE4/APOE4	184 (13.8%)	20 (1.4%)	204 (7.5%)	114 (8.5%)	14 (1.2%)	128 (5.1%)	332 (6.4%)

DISCUSSION

We developed a novel computational framework to explore genetically regulated pathways PRS (**GRPa-PRS**) among predefined risk strata groups in polygenic diseases. We used AD as a prototype and examined two independent cohorts to cross-validate our framework and findings. Given the dominant effect of *APOE*, we aimed to identify the GRPa-PRS among different AD risk strata which are independent of the dominant factor of *APOE.* Moreover, we disentangled the GRPa-PRS that is orthogonal to, not just correlated to, their AD risks, which unveils novel resilience and susceptibility factors underlying the risk strata group. Notably, we identified one significant resilience GRPa (MF 'thiolester hydrolase activity') replicated in comparison from two different PRS models and strata groups. Lastly, we benchmarked our GRPa-PRS with one pathway PRS method, and we verified that our framework possesses more variance explained and more pathways independent of the *APOE* effect.

Novel risk strata assumption and parallel design

There are a few assumptions for the conjecture's construction before we could define our risk strata. 1) the diagnosis of each individual mostly reflect their ultimate AD status; 2) the individual AD genetic burden is largely reflected by the PRS; 3) the individuals in each PRS stratified

group might carry different AD genetic risk, but the overall genetic burden is similar; 4) the individuals are stratified to extreme quantiles (top and bottom 15%) to reach a maximum PRS disparity as well as a balanced minimum individuals to reach statistical power; 5) considerable proportion of common variant genetic variation can be captured by the variation of GReX genes and their pathway. In other words, the genetic risk difference between subgroups could be eventually reflected by the GReX difference and functional pathway difference.

With these conjectures, we theorized that different strata, as defined by extremes of PRS load, would provide us with hints as to different resilience and extra burden or susceptibility genetic factors by identifying those who defy genetic odds by remaining CN despite a high AD genetic load and those who develop AD despite a low genetic load, and also those who conform to genetic odds by remaining CN in concordance to a low AD genetic load and those with who develop AD in concordance to a high genetic load. By systematically contrasting these different subgroups, our subgroup design allowed us to innovatively ascertain genetic factors related to resilience and extra burden in our exemplar, AD. The identification of such genetic factors would be a first step towards a general framework that would enable preventative efforts at large by the enhancement of resilience factors and the inhibition of the susceptibility factors. Additionally, by linking genetics to genetically-regulated expression (GReX) using TWAS and identifying genes that are protein coding, we made our newly developed subgroup design and **GRPa-PRS** framework potentially actionable.

To ensure the robustness of our findings, we developed two approaches (GRPa-GSEA and GRPa-GSVA) to systematically characterize the GRPas among risk strata predefined by three GWAS summary statistics and two independent cohorts. Overall, similar trends of GRPa intensity among parallel groups were observed (**Fig. 3 and Fig. 4**). However, due to the slight difference in each group, the significant pathways might not be always shared. As shown in **Fig. 5B**, well-known AD-related pathways (including amyloid-beta clearance, tau protein binding, and

astrocytes response to oxidative stress) identified by these two approaches are highlighted consistently in case_control, TBall, and TBAD groups in Model 1, suggesting stable GRPas could be observed. However, in Model 2 results, GRPa-GSVA identified more immune-related functions (**Fig. 3D, Fig. S2C, Fig. S2D**), while GRPa-GSEA identified more metabolism- and neuron- related functions (**Fig. 4C, Fig. 4D, Fig. S4C, Fig. S4D**), which suggests the *APOE*-independent GRPas have smaller effect sizes than the *APOE*-related GRPas. This could also explain we identified more GRPas TB15AD (*APOE*-dependent GRPas) group than TB15Ctr group.

Resilience-related and susceptibility-related pathways, molecular functions, and cellular components

There is no effective drug for AD yet, therefore, novel pathways that could mitigate or intervene in AD pathogenesis are in urgent need. Through our GRPa-PRS framework, we could identify novel underlying genetically regulated pathways that contribute to resilience. We systematically defined new resilience-related GRPas with two major features: 1) Those pathways and terms should be identified from TBCtr or T groups of Model 2, which are independent of the *APOE*-related functions. To be notified, these GRPas might exist in Model 1 or other comparison groups as well. 2) Since the resilience effects should be orthogonal and not just opposite to PRS risks, correlation tests should be conducted and non-significant to verify the orthogonality.

We observed three verified resilience-related and susceptibility-related gene sets, as well as three potential resilience-related gene sets. Firstly, our major findings in GO term explicitly related to resilience was 'thioester hydrolase activity' (**Fig. 4C and Fig. 6A**), which is a molecular function (MF) term composed of the activity of several enzymes known as thioester hydrolases or thioesterases; this large enzyme group hydrolyzes the thioester bond between a carbonyl group and a sulfur atom; most of the Acyl-Coenzyme A (Acyl-CoA) thioesterases act

specifically on molecular substrates that contain coenzyme A [66,67]. Acyl-CoA thioesterases catalyze the hydrolysis of acyl-CoAs to the free fatty acid and coenzyme A, having the capacity to regulate intracellular levels of acyl-CoAs, free fatty acids and coenzyme A. These enzymes are ubiquitously located in most cellular compartments such as cytosol, mitochondria, and peroxisomes [68]. In our analyses, we found 34 genes enriched in this specific MF, including 10 *ACOT* (Acyl-CoA Thioesterase) genes. *ACOT*s, most notably, hydrolyze Acyl-CoA with the group of enzymes Acyl-CoA thioesterases. Specifically, the cytosolic enzyme acyl coenzyme A thioester hydrolase (Acot7), coded by the *ACOT7* gene, also known as *BACH* (brain acyl-CoA hydrolase) gene, is abundantly expressed and active mainly in brain tissue, particularly neurons, and is identified as an enzyme responsible for the long-chain acyl-CoA hydrolyzing activity in the brain [68]. *ACOT7* expression is induced during embryogenesis in association with neuronal differentiation and continues in neurons at postnatal stages, resulting in high expression of *ACOT7* in the adult brain in neurons [67]. It has become increasingly evident that many neurodegenerative diseases, including AD, have metabolic dysfunction that often involves dysregulation in lipid metabolism. Fatty acid metabolism in the nervous system has been shown to regulate physiology and behavior directly. Accordingly, *ACOT7* knockout mice show neurodegeneration, neurological, and behavioral deficits, and it has been suggested that *ACOT7* is required to maintain neuronal fatty acid homeostasis and that dysregulation of *ACOT7* results in neural lipotoxicity; thus, *ACOT7* enzyme expression would have a neuroprotective effect, being an important regulator in neurometabolism that protects neurons against the damaging effects of fatty acid excess [69]. ACOTs also act as enzymes in the α- and β-oxidation of various lipids in peroxisomes. Human peroxisomes contain at least two ACOTs (from the genes *ACOT4* and *ACOT8*), both of which were found to be enriched in our analyses. *ACOT8* shows a broad tissue expression, and expression is regulated by fasting via the peroxisome proliferator-activated receptor alpha (PPARα), hinting at a role for ACOT8 in the regulation of fatty acid oxidation [70]. Interestingly, metformin, a drug that is a PPARα agonist,

has been associated with a decreased risk of dementia [71,72]. The Acot11 enzyme, expressed by the gene *ACOT11*, also found enriched in our analyses, is specific toward medium- and long-chain acyl-CoA molecules, and may be involved in obesity, a known modifiable risk factor for AD [66,73,74].

Another MF that appeared related to resilience, however only when comparing the top 15% vs. bottom 15% of controls, was 'carbohydrate transmembrane transporter activity' (**Fig. 4C and Fig. 6B**) . Broadly speaking, this MF enables transmembrane transfer of carbohydrates. Specifically, many of the genes that we found enriched in our analyses regarding this pathway are related to glucose metabolism and its transport in general. Among others, the gene Solute Carrier Family 2 Member 1 (*SLC2A1*, also known as *GLUT-1*), encodes the protein 'solute carrier family 2, facilitated glucose transporter member 1', which is a facilitative glucose transporter responsible for basal glucose uptake [75–79]. Importantly, this protein is the most important energy carrier in the brain, being present at the blood-brain barrier and assuring the facilitative transport of glucose into the brain in an energy-independent manner [79]. Another gene that we found enriched, pertaining to the same family, is Solute Carrier Family 2 Member 4 (*SLC2A4* or *GLTU-4*). This gene encodes the protein Solute carrier family 2, facilitated glucose transporter member 4, or Glucose transporter type 4, insulin-responsive protein, which regulates the removal of glucose from circulation into muscle and adipose cells. Another one is the gene *SLC17A5* (Solute Carrier Family 17 Member 5), which regulates lysosome function and metabolism of sialylated conjugates that impact oligodendrocyte lineage differentiation and myelinogenesis in the brain [80–83]. It might also participate in the synaptic vesicular storage of glutamate, being important to the aspartegic-associated glutamatergic neurotransmission and activation of metabotropic receptors that inhibit subsequent transmitter release [81–83].

The canonical pathway 'histidine metabolism' (**Fig. S5C, Fig. S9A, and Fig. S9B**), contrary to 'carbohydrate transmembrane transporter activity', was enriched in the top 15% of controls

compared to the top 15% of cases, thus being also enriched in a comparison directly related to resilience. The metabolism of histidine yields glutamate; it can also form carnosine, an abundant peptide that acts as an antioxidant in the brain [90]. In an animal model fed with a high fat diet, carnosine was able to revert the decline in contextual memory and prevented cognitive decline. Carnosine also decreased microglial activation in this model [91].

From the GRPa-GSEA, we also identified three significant gene sets in TB15Ctr or T15 groups. As these terms were not significant in GRPa-GSVA approaches, therefore, their orthogonality warrant further investigation. One CC which also was found enriched in the top 15% of controls when compared to the top 15% of cases is 'gamma secretase complex' (**Fig. S2D**). The gamma-secretase complex is a protein complex whose endopeptidase activity contains a presenilin catalytic subunit, either *PSEN1* or *PSEN2*, as the core complex [84,85]. Gamma-secretase catalyzes the intramembrane proteolysis of Notch, the β-amyloid precursor protein, as a key step in the pathogenesis of AD [86].

Regarding the bottom 15% of cases and bottom 15% of controls, a comparison related to extra-burden, the CC found altered is the 'dense core granule' (**Fig. 3B**). These are organelles located in neurons and neuroendocrine cells, which are specialized secretory cells, and are related to transport and exocytosis [87]. One study which performed a proteomic profile in the brains of people with AD found that several proteins involved in exocytosis show altered levels in AD [88]. Another study found that early synaptic dysfunctions in AD are composed of alterations in the presynaptic vesicle exocytosis of glutamate [89].

The last term, we identified is another canonical pathway 'formation of apoptosome' (**Fig. S3D**), Alzheimer's disease (AD) is characterized by the accumulation of hyperphosphorylated tau and amyloid-β (Aβ) protein resulting in synaptic loss and apoptosis. Aβ and tau deposition trigger apoptotic pathways that result in neuronal death [92]. Aβ was reported to bind procaspase-9 to

inhibit the assembly of Apaf-1 apoptosome in mice and therefore induce the intrinsic mitochondria-dependent apoptotic pathway [93]. Therefore, Apoptosis is considered to be responsible for manifestations associated with AD under pathological conditions.

Despite the six resilience-related pathways and terms, we only identified one significant susceptibility-related term CC 'Inhibitory synapse' in **Fig. S4C** KB15 group (**Fig. S9C and Fig. S9D**). The loss of excitatory synapses in Alzheimer's disease (AD) is well recognized. The pathological accumulation of Aβ contributes significantly to the dysfunction and depletion of excitatory synapses [94]. And the synaptic disturbances in excitatory to inhibitory (E/I) balance have been observed in AD patients [95]. However, direct evidence for how genetically disturbed inhibitory synapse function could bring extra burden to the initiation of AD pathogenesis remains unclear. Overall, our proposed framework can successfully identify specific meaningful resilience- and susceptibility-related genes, terms, and pathways.

Limitations

Our work is not without limitations. First, due to a lack of individual information in the GWAS summary statistics, the overlapped individuals among our two cohorts and GWAS summary statistics could not be detected. This might have led to bias in the PRS calculation. Given the large samples (at most less 1% of samples will be overlapped) in GWAS summary statistics, we believed this impact of the possible bias to be relatively small. Second, by borrowing the expression information from diseases/traits-related tissues, our GRPa-PRS framework has more power to detect GRPa independent from *APOE* at the cost of higher computation burden and time than SNP-based methods like PRSet. Third, although we identified relatively consistent signals in Model 1 (TBall, TBAD, and TBCtr) among three GWAS summary statistics and the two independent cohorts using our two approaches in the GRPa-PRS framework, we failed to identify relatively stable signals in Model 2 and other groups (T and B) in either Model 1

or Model 2. One potential explanation for this is the small sample size and small effect size of resilience GRPas compared to APOE-related GRPas, as expected. Fourth, our newly developed method does not consider the effect of rare variants; this fact might have had a larger impact in the comparisons involving the susceptible group where rare variants possibly play a larger role due to their higher susceptibility. Fifth, our newly developed framework relies on good PRS performance, which in turn depends on the specific polygenicity and heritability of the considered disorder. AD has moderate polygenicity, with 38 loci identified by Wightman et al. [92], and SNP-based heritability estimated to be as high as 66% [96], thus GRPa-PRS should have shown at least a relatively good performance. However, the high polygenicity and heritability requirements limit the generalization of our model. Sixth, here we choose a soft threshold of 15% and meet a minimum sample size of 50 in each comparison extreme subgroup and simultaneously maximize the difference. Notably, the results of soft thresholds of 10% and 20% are comparable to 15%. We acknowledged we would have more power to detect more significant GRPas, if a large sample size is available. Therefore, when applying our model to other polygenic diseases using larger cohorts, a more stringent threshold to define the subgroups would maximize their genetic differences. Seventh, we also observe that it is a tendency for more APOE4 alleles in the high PRS group than in the low PRS group. We design to conduct the differential GRPas between groups without using APOE4 allele dosage as the covariate given the dosage difference listed in **Table S1-S6**. Therefore, we will not directly observe if any GRPa could be significant after accounting for APOE4 allele dosage. However, we believe this result is addressed by our Model 2 design. Finally, the minimum age of the included controls was 57 years, thus we cannot guarantee that some individuals belonging to the resilient group will not eventually develop AD as they age. However, we believe that this makes our model conservative when searching for resilience factors and that it leads to false-negatives but increased the odds that the found signals are indeed true-positives. In addition,

age of onset as a limitation applies mostly to our exemplar AD and should play a considerably lesser role in other polygenic disorders with development earlier in life.

Future directions

Here we used AD as a prototype. We plan to extend and validate GRPa-PRS to other polygenic disorders which possess good PRS performance, as is the case of some psychiatric disorders such as depression and schizophrenia. Such disorders do not possess a dominant gene as is the case of *APOE* and thus would be more straightforward to implement and interpret. Although here we focused on the identification of resilience factors, GRPa-PRS using our subgrouping framework can equally be used to identify extra burden or susceptibility factors by focusing on the groups related to extra burden. Another exciting possibility is to incorporate into GRPa-PRS the capacity of identifying different subgroups with different resilience and susceptibility factors; this could enable personalized prevention by reinforcing the different resilience factors and weakening the different susceptibility factors of each individual according to their specific PRS subgroup instead of reinforcing or mitigating them universally in all individuals. Thus, the identification of cluster pathways using GRPa-PRS could guide precise prevention. Finally, the incorporation of the disease strata according to PRS developed in GRPa-PRS to single-cell data could lead to new biological insights not possible using PRS alone.

Conclusions

We developed a framework, **GRPa-PRS**, including two approaches (GRPa-GSEA and GRPa-GSVA) to systematically explore the differentially genetically-regulated pathways among individuals stratified by their AD risks estimated by PRS, ultimately revealing resilience and susceptibility biological factors. The GReX-level comparison among those groups unveiled pathways associated with AD risks, including many well-known pathways as well as new

pathways. We highlighted one metabolic pathway 'thiolester hydrolase activity' playing resilient role in two resilient group comparisons. This GRPa-based PRS framework can be extended to other polygenic complex disorders and can identify unknown counterfactors, shedding light to novel risk and preventive pathways.

DATA AVAILABILITY

All the data generated or analyzed in this study is available from the authors upon reasonable request. The overall framework can be downloaded from GitHub.

ACKNOWLEDGEMENTS

We would like to thank all the members of Bioinformatics and Systems Medicine Laboratory (BSML) for constructive discussion.

FUNDING

This research was partially supported by National Institutes of Health grants awarded to Z.Z. (R01LM012806, R01DE030122, and R01DE029818). We thanked the resource support from Cancer Prevention and Research Institute of Texas (CPRIT RP180734 and RP210045). A.L. is supported by a training fellowship from the Gulf Coast Consortia on Training in Precision Environmental Health Sciences (TPEHS) Training Grant (T32ES027801).

REFERENCES

1. Sims R, Hill M, Williams J. The multiplex model of the genetics of Alzheimer's disease. Nat Neurosci. 2020;23:311–22.

2. Leonenko G, Baker E, Stevenson-Hoare J, Sierksma A, Fiers M, Williams J, et al. Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. Nat Commun. 2021;12:4506.

3. de Rojas I, Moreno-Grau S, Tesi N, Grenier-Boley B, Andrade V, Jansen IE, et al. Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. Nat Commun. 2021;12:3417.

4. Escott-Price V, Sims R, Bannister C, Harold D, Vronskaya M, Majounie E, et al. Common polygenic variation enhances risk prediction for Alzheimer's disease. Brain. 2015;138:3673–84.

5. Escott-Price V, Myers AJ, Huentelman M, Hardy J. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. Ann Neurol. 2017;82:311–4.

6. Baker E, Escott-Price V. Polygenic Risk Scores in Alzheimer's Disease: Current Applications and Future Directions. Front Digit Health. 2020;2:14.

7. 2022 Alzheimer's disease facts and figures. Alzheimers Dement. 2022;18:700–89.

8. Mendez MF. Early-onset Alzheimer Disease and Its Variants. Continuum . 2019;25:34–51.

9. Lanoiselée H-M, Nicolas G, Wallon D, Rovelet-Lecrux A, Lacour M, Rousseau S, et al. APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases. PLoS Med. 2017;14:e1002270.

10. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet. 2019;51:404–13.

11. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat Genet. 2019;51:414–30.

12. Wightman DP, Jansen IE, Savage JE, Shadrin AA, Bahrami S, Holland D, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nat Genet. 2021;53:1276–82.

13. Schwartzentruber J, Cooper S, Liu JZ, Barrio-Hernandez I, Bello E, Kumasaka N, et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. Nat Genet. 2021;53:392–402.

14. Holstege H, Hulsman M, Charbonnier C, Grenier-Boley B, Quenez O, Grozeva D, et al. Exome sequencing identifies rare damaging variants in ATP8B4 and ABCA1 as risk factors for Alzheimer's disease. Nat Genet. 2022;54:1786–94.

15. Shigemizu D, Asanomi Y, Akiyama S, Mitsumori R, Niida S, Ozaki K. Whole-genome sequencing reveals novel ethnicity-specific rare variants associated with Alzheimer's disease. Mol Psychiatry. 2022;27:2554–62.

16. Ming C, Wang M, Wang Q, Neff R, Wang E, Shen Q, et al. Whole genome sequencing-based copy number variations reveal novel pathways and targets in Alzheimer's disease. Alzheimers Dement. 2022;18:1846–67.

17. Saddiki H, Fayosse A, Cognat E, Sabia S, Engelborghs S, Wallon D, et al. Age and the association between apolipoprotein E genotype and Alzheimer disease: A cerebrospinal fluid biomarker-based case-control study. PLoS Med. 2020;17:e1003289.

18. Dumitrescu L, Mahoney ER, Mukherjee S, Lee ML, Bush WS, Engelman CD, et al. Genetic variants and functional pathways associated with resilience to Alzheimer's disease. Brain. 2020;143:2561–75.

19. Liu A, Manuel AM, Dai Y, Fernandes BS, Enduru N, Jia P, et al. Identifying candidate genes and drug targets for Alzheimer's disease by an integrative network approach using genetic and brain region-specific proteomic data. Hum Mol Genet. Oxford University Press (OUP); 2022;31:3341–54.

20. Tesi N, van der Lee SJ, Hulsman M, Jansen IE, Stringa N, van Schoor NM, et al. Immune response and endocytosis pathways are associated with the resilience against Alzheimer's disease. Transl Psychiatry. 2020;10:332.

21. Stern Y. Cognitive reserve in ageing and Alzheimer's disease. Lancet Neurol. 2012;11:1006–12.

22. Hess JL, Tylee DS, Mattheisen M, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Børglum AD, et al. A polygenic resilience score moderates the genetic risk for schizophrenia. Mol Psychiatry. 2021;26:800–15.

23. Hou J, Hess JL, Armstrong N, Bis JC, Grenier-Boley B, Karlsson IK, et al. Polygenic resilience scores capture protective genetic effects for Alzheimer's disease. Transl Psychiatry. 2022;12:296.

24. Escott-Price V, Schmidt KM. Probability of Alzheimer's disease based on common and rare genetic variants. Alzheimers Res Ther. 2021;13:140.

25. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9:1825.

26. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019;51:592–9.

27. Dai Y, Wang J, Jeong H-H, Chen W, Jia P, Zhao Z. Association of CXCR6 with COVID-19 severity: delineating the host genetic factors in transcriptomic regulation. Hum Genet. 2021;140:1313–28.

28. Dai Y, Jia P, Zhao Z, Gottlieb A. A method for bridging population-specific genotypes to detect gene modules associated with Alzheimer's disease. Cells. MDPI AG; 2022;11:2219.

29. Dai Y, Pei G, Zhao Z, Jia P. A Convergent Study of Genetic Variants Associated With

Crohn's Disease: Evidence From GWAS, Gene Expression, Methylation, eQTL and TWAS. Front Genet. 2019;10:318.

30. Chatzinakos C, Georgiadis F, Lee D, Cai N, Vladimirov VI, Docherty A, et al. TWAS pathway method greatly enhances the number of leads for uncovering the molecular underpinnings of psychiatric disorders. Am J Med Genet B Neuropsychiatr Genet. 2020;183:454–63.

31. Lee JH, Cheng R, Graff-Radford N, Foroud T, Mayeux R, National Institute on Aging Late-Onset Alzheimer's Disease Family Study Group. Analyses of the National Institute on Aging Late-Onset Alzheimer's Disease Family Study: implication of additional loci. Arch Neurol. 2008;65:1518–26.

32. Li H, Wetten S, Li L, St Jean PL, Upmanyu R, Surh L, et al. Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. Arch Neurol. 2008;65:45–53.

33. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious Orders Study and Rush Memory and Aging Project. J Alzheimers Dis. 2018;64:S161–89.

34. Allen M, Carrasquillo MM, Funk C, Heavner BD, Zou F, Younkin CS, et al. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Sci Data. 2016;3:160089.

35. Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci Data. 2018;5:180185.

36. Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, et al. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology. 2010;74:201–9.

37. Cammann D, Lu Y, Cummings MJ, Zhang ML, Cue JM, Do J, et al. Genetic correlations between gut microbiome genera, Alzheimer's disease diagnosis, and APOE genotypes: a polygenic risk score study [Internet]. 2022. Available from: https://www.researchsquare.com/article/rs-2292371/v1

38. McCarthy Tools [Internet]. [cited 2023 Mar 12]. Available from: https://www.well.ox.ac.uk/~wrayner/tools/

39. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284–7.

40. De Jager PL, Shulman JM, Chibnik LB, Keenan BT, Raj T, Wilson RS, et al. A genome-wide scan for common variants affecting the rate of age-related cognitive decline. Neurobiol Aging. 2012;33:1017.e1-15.

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

42. Wang Q, Chen K, Su Y, Reiman EM, Dudley JT, Readhead B. Deep learning-based brain

transcriptomic signatures associated with the neuropathological and clinical severity of Alzheimer's disease. Brain Commun. 2022;4:fcab293.

43. Danecek P, Auton A, Abecasis G, Albers CA, Banks E. Sherry ST et al. 2011. The variant call format and VCFtools. Bioinformatics.

44. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

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

46. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience [Internet]. 2021;10. Available from: http://dx.doi.org/10.1093/gigascience/giab008

47. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26:2867–73.

48. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.

49. PLINK: Whole genome data analysis toolset [Internet]. [cited 2023 Mar 13]. Available from: https://zzz.bwh.harvard.edu/plink/index.shtml

50. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467:52–8.

51. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. Bioinformatics. 2020;36:5424–31.

52. Privé F, Aschard H, Ziyatdinov A, Blum MGB. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics. 2018;34:2781–7.

53. Neff RA, Wang M, Vatansever S, Guo L, Ming C, Wang Q, et al. Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. Sci Adv [Internet]. 2021;7. Available from: http://dx.doi.org/10.1126/sciadv.abb5398

54. Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci U S A. 2004;101:2173–8.

55. Blalock EM, Buechel HM, Popovic J, Geddes JW, Landfield PW. Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease. J Chem Neuroanat. 2011;42:118–26.

56. Goudriaan A, de Leeuw C, Ripke S, Hultman CM, Sklar P, Sullivan PF, et al. Specific glial functions contribute to schizophrenia susceptibility. Schizophr Bull. 2014;40:925–35.

57. Ruano D, Abecasis GR, Glaser B, Lips ES, Cornelisse LN, de Jong APH, et al. Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability. Am J

Hum Genet. 2010;86:113–25.

58. Lips ES, Cornelisse LN, Toonen RF, Min JL, Hultman CM, International Schizophrenia Consortium, et al. Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia. Mol Psychiatry. 2012;17:996–1006.

59. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. 2019;47:W199–205.

60. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1:417–25.

61. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data [Internet]. Nature Genetics. 2015. p. 1091–8. Available from: http://dx.doi.org/10.1038/ng.3367

62. Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genet. 2019;15:e1007889.

63. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.

64. Choi SW, García-González J, Ruan Y, Wu HM, Porras C, Johnson J, et al. PRSet: Pathway-based polygenic risk score analyses and software. PLoS Genet. 2023;19:e1010624.

65. Levitis DA, Lidicker WZ, Freund G. Behavioural biologists don't agree on what constitutes behaviour. Anim Behav. Elsevier BV; 2009;78:103–10.

66. Cantu DC, Chen Y, Reilly PJ. Thioesterases: a new perspective based on their primary and tertiary structures. Protein Sci. 2010;19:1281–95.

67. Yamada J. Long-chain acyl-CoA hydrolase in the brain. Amino Acids. 2005;28:273–8.

68. Hunt MC, Alexson SEH. The role Acyl-CoA thioesterases play in mediating intracellular lipid metabolism. Prog Lipid Res. 2002;41:99–130.

69. Ellis JM, Wong GW, Wolfgang MJ. Acyl coenzyme A thioesterase 7 regulates neuronal fatty acid metabolism to prevent neurotoxicity. Mol Cell Biol. 2013;33:1869–82.

70. Hunt MC, Siponen MI, Alexson SEH. The emerging role of acyl-CoA thioesterases and acyltransferases in regulating peroxisomal lipid metabolism. Biochim Biophys Acta. 2012;1822:1397–410.

71. Campbell JM, Stephenson MD, de Courten B, Chapman I, Bellman SM, Aromataris E. Metformin Use Associated with Reduced Risk of Dementia in Patients with Diabetes: A Systematic Review and Meta-Analysis. J Alzheimers Dis. 2018;65:1225–36.

72. Liao W, Xu J, Li B, Ruan Y, Li T, Liu J. Deciphering the Roles of Metformin in Alzheimer's Disease: A Snapshot. Front Pharmacol. 2021;12:728315.

73. Flores-Cordero JA, Pérez-Pérez A, Jiménez-Cortegana C, Alba G, Flores-Barragán A, Sánchez-Margalet V. Obesity as a Risk Factor for Dementia and Alzheimer's Disease: The Role

of Leptin. Int J Mol Sci [Internet]. 2022;23. Available from: http://dx.doi.org/10.3390/ijms23095202

74. Slomski A. Obesity Is Now the Top Modifiable Dementia Risk Factor in the US. JAMA. 2022;328:10.

75. Mueckler M, Makepeace C. Transmembrane segment 6 of the Glut1 glucose transporter is an outer helix and contains amino acid side chains essential for transport activity. J Biol Chem. Elsevier BV; 2008;283:11550–5.

76. Mueckler M, Makepeace C. Model of the exofacial substrate-binding site and helical folding of the human Glut1 glucose transporter based on scanning mutagenesis. Biochemistry. American Chemical Society (ACS); 2009;48:5934–42.

77. Lee EE, Ma J, Sacharidou A, Mi W, Salato VK, Nguyen N, et al. A protein kinase C phosphorylation motif in GLUT1 affects glucose transport and is mutated in GLUT1 deficiency syndrome. Mol Cell. Elsevier BV; 2015;58:845–53.

78. Kapoor K, Finer-Moore JS, Pedersen BP, Caboni L, Waight A, Hillig RC, et al. Mechanism of inhibition of human glucose transporter GLUT1 is conserved between cytochalasin B and phenylalanine amides. Proc Natl Acad Sci U S A. Proceedings of the National Academy of Sciences; 2016;113:4711–6.

79. Klepper J, Wang D, Fischbarg J, Vera JC, Jarjour IT, O'Driscoll KR, et al. Defective glucose transport across brain tissue barriers: a newly recognized neurological syndrome. Neurochem Res. 1999;24:587–94.

80. Morin P, Sagné C, Gasnier B. Functional characterization of wild-type and mutant human sialin. EMBO J. Wiley; 2004;23:4560–70.

81. Miyaji T, Omote H, Moriyama Y. Functional characterization of vesicular excitatory amino acid transport by human sialin. J Neurochem. Wiley; 2011;119:1–5.

82. Qin L, Liu X, Sun Q, Fan Z, Xia D, Ding G, et al. Sialin (SLC17A5) functions as a nitrate transporter in the plasma membrane. Proc Natl Acad Sci U S A. Proceedings of the National Academy of Sciences; 2012;109:13434–9.

83. Lodder-Gadaczek J, Gieselmann V, Eckhardt M. Vesicular uptake of N-acetylaspartylglutamate is catalysed by sialin (SLC17A5). Biochem J. Portland Press Ltd.; 2013;454:31–8.

84. Shirotani K, Edbauer D, Prokop S, Haass C, Steiner H. Identification of distinct gamma-secretase complexes with different APH-1 variants. J Biol Chem. Elsevier BV; 2004;279:41340–5.

85. Zhou S, Zhou H, Walian PJ, Jap BK. CD147 is a regulatory subunit of the gamma-secretase complex in Alzheimer's disease amyloid beta-peptide production. Proc Natl Acad Sci U S A. Proceedings of the National Academy of Sciences; 2005;102:7499–504.

86. Kimberly WT, LaVoie MJ, Ostaszewski BL, Ye W, Wolfe MS, Selkoe DJ. Gamma-secretase is a membrane protein complex comprised of presenilin, nicastrin, Aph-1, and Pen-2. Proc Natl Acad Sci U S A. Proceedings of the National Academy of Sciences; 2003;100:6382–7.

87. Silverman MA, Johnson S, Gurkins D, Farmer M, Lochner JE, Rosa P, et al. Mechanisms of transport and exocytosis of dense-core granules containing tissue plasminogen activator in developing hippocampal neurons. J Neurosci. Society for Neuroscience; 2005;25:3095–106.

88. Musunuri S, Khoonsari PE, Mikus M, Wetterhall M, Häggmark-Mänberg A, Lannfelt L, et al. Increased levels of extracellular microvesicle markers and decreased levels of endocytic/exocytic proteins in the Alzheimer's disease brain. J Alzheimers Dis. IOS Press; 2016;54:1671–86.

89. Perdigão C, Barata MA, Araújo MN, Mirfakhar FS, Castanheira J, Guimas Almeida C. Intracellular trafficking mechanisms of synaptic dysfunction in Alzheimer's disease. Front Cell Neurosci. Frontiers Media SA; 2020;14:72.

90. Morris ML, Lee SC, Harper AE. Influence of differential induction of histidine catabolic enzymes on histidine degradation in vivo. J Biol Chem. Elsevier BV; 1972;247:5793–804.

91. Herculano B, Tamura M, Ohba A, Shimatani M, Kutsuna N, Hisatsune T. β-alanyl-L-histidine rescues cognitive deficits caused by feeding a high fat diet in a transgenic mouse model of Alzheimer's disease. J Alzheimers Dis. IOS Press; 2013;33:983–97.

92. Kumari S, Dhapola R, Reddy DH. Apoptosis in Alzheimer's disease: insight into the signaling pathways and therapeutic avenues. Apoptosis [Internet]. Springer Science and Business Media LLC; 2023; Available from: http://dx.doi.org/10.1007/s10495-023-01848-y

93. Sharoar MG, Islam MI, Shahnawaz M, Shin SY, Park I-S. Amyloid β binds procaspase-9 to inhibit assembly of Apaf-1 apoptosome and intrinsic apoptosis pathway. Biochim Biophys Acta. Elsevier BV; 2014;1843:685–93.

94. Palop JJ, Mucke L. Amyloid-beta-induced neuronal dysfunction in Alzheimer's disease: from synapses toward neural networks. Nat Neurosci. Springer Science and Business Media LLC; 2010;13:812–8.

95. Lauterborn JC, Scaduto P, Cox CD, Schulmann A, Lynch G, Gall CM, et al. Increased excitatory to inhibitory synaptic ratio in parietal cortex samples from individuals with Alzheimer's disease. Nat Commun. Springer Science and Business Media LLC; 2021;12:2603.

96. Baker E, Leonenko G, Schmidt KM, Hill M, Myers AJ, Shoai M, et al. What does heritability of Alzheimer's disease represent? PLoS One. 2023;18:e0281440.

FIGURE LEGENDS

Fig. 1. GRPa Workflow and subgrouping strata. The blue tabulates in the top row indicate the input data from preprocessed individual genotyping data, AD GWAS summary statistics, curated gene sets. The green tabulates and background include the key steps of our GRPa-PRS framework: 1. PRS strata strategy; 2. GReX imputation; 3.a GRPa-GSEA approach and 3.b GRPa-GSVA; 4.a Differential GRPas summary and 4.b Orthogonality test. The dark blue tabulate includes the method PRSet we used to benchmark the performance of differential GRPa. The dash line indicates the comparison between three approaches, PRSet, GRPa-GSEA, and GRPa-GSVA. The orange tabulates indicate designed to explore the differential GRPa from three different approaches and their orthogonality. The highlighted two groups in Step 1: Resilient group (high-risk controls) and Susceptible group (low-risk cases) defined in Table 1.

Fig. 2. Constructing the optimum PRS model using AD GWAS summary statistics. (A) Subgroup cohorts by PRS models to evaluate GRPa in different risk groups as shown in Table 1: (1) case_control, (2) TB15all, (3) TB15AD, (4) TB15Ctr, (5) T15, (6) B15.
In the process of constructing the optimum PRS, we choose the best model according to maximum AUC for (B) discovery cohort and (C) replication cohort. These two scatter plots show how AUC changes with tuning hyperparameters, including p (probability of causal variant), h^2 (heritability), and sparse (shrinks part of coefficients to 0). (D) shows how AUC changes with different p-value thresholds.

Fig. 3. GRPa-GSEA results enrich on GO pathways. GRPa identified by GRPa-GSEA under Model 1, the model using full genotype to detect all pathways associated with subgroup comparison, and Model 2, the model using excluding *APOE* region genotype to detect pathways

associated with subgroup comparison and independent with *APOE* effect. (A) GO GRPa identified in discovery dataset Model 1 (full model), (B) GO GRPa identified in replication dataset Model 1 (full model), (C) GO GRPa identified in discovery dataset Model 2 (no-APOE model), no significant result (FDR < 0.05) in this condition, and (D) GO GRPa identified in discovery dataset Model 2 (no-APOE model). The Venn plot for overlapping signals between the discovery cohort and replication cohort under Model 1 was shown in (E). * The *x* axis shows the heatmap list of the subgroup comparison based on different GWAS summary statistics. S represents Schwartzentruber et al; K represents Kunkle et al; W represents Wightman et al.

Fig. 4. GRPa-GSVA results enrich GO pathways. GRPas identified by GRPa-GSVA under Model 1, the model using full genotype to detect all pathways associated with subgroup comparison, and Model 2, the model using excluding *APOE* region genotype to detect pathways associated with subgroup comparison and independent with *APOE* effect. (A) GO GRPa identified in discovery dataset Model 1 (full model), (B) GO GRPa identified in replication dataset Model 1 (full model), (C) GO GRPa identified in discovery dataset Model 2 (no-APOE model) , no significant results in this condition, and (D) GO GRPa identified in discovery dataset Model 2 (no-APOE model). The Venn plot for overlapping signals between discovery cohort and replication cohort under Model 1 were shown in (E). * The *x* axis shows the heatmap list of the subgroup comparison based on different GWAS summary statistics. S represents Schwartzentruber et al; K represents Kunkle et al; W represents Wightman et al.

Fig. 5. Compare results of three methods GRPa-GSEA, GRPa-GSVA, PRSet in GReX with *APOE*. The overlapping signals between GRPa-GSEA, GRPa-GSVA, PRSet were shown in (A), including comparison between cases vs control, comparison between individuals with top PRS vs individuals with bottom PRS, comparison between AD cases with top PRS vs AD cases with bottom PRS. We compare the combination of signals from discovery cohort and replication

cohort between GRPa-GSEA, GRPa-GSVA and listed their overlapping signals in (B). (C) shows the R2 improvement comparison between formula 6 and formula 8.

Fig. 6. Further exploration for identified pathways. We visualize the correlation between Sprs and (A) thiolester hydrolase activity, and (C) carbohydrate transmembrane transporter activity, (E) amyloid beta clearance respectively. In addition, we plot the distribution difference within STB15all, STB15Ctr, WTB15Ctr, KT15 for (B) thiolester hydrolase activity, and (D) carbohydrate transmembrane transporter activity, (F) amyloid beta clearance respectively to validate heatmap results.

Abbreviation

AD: Alzheimer's disease; AUC: area under the receiver operating curve; BP: Biological Process; CC: Cellular Component; CI: confidence interval; CN: cognitively normal; CDR: clinical dementia rating; FDR: false discovery rate; eQTL: expression quantitative trait loci; GO: gene ontology; GRPa: genetically-regulated pathway; GRPa-GSEA: genetically-regulated pathway-gene-set enrichment analysis; GRPa-GSVA: genetically-regulated pathway-gene-set variational analysis; GReX: genetically-regulated expression; GSEA: gene-set enrichment analysis; GSVA: gene-set variational analysis; GWAS: genome-wide association study; GWAX: genome-wide association study based on proxy phenotype; h^2: heritability; KEGG: Kyoto Encyclopedia of Genes and Genomes; LOAD: late-onset Alzheimer’s Disease; LogisticLRT: logistic likelihood ratio test; MF: Molecular Function; OR: odds ratio; PRS: polygenic risk scores; TWAS: transcriptome-wide association study; WGS: Whole Genome Sequencing;

Subgroup:

T: top; B: bottom; ; TB, top vs bottom; Ctr: control;

TBall: top vs bottom for all individuals; TBAD: top vs bottom for only AD patients; TBCtr: top vs bottom for only controls;

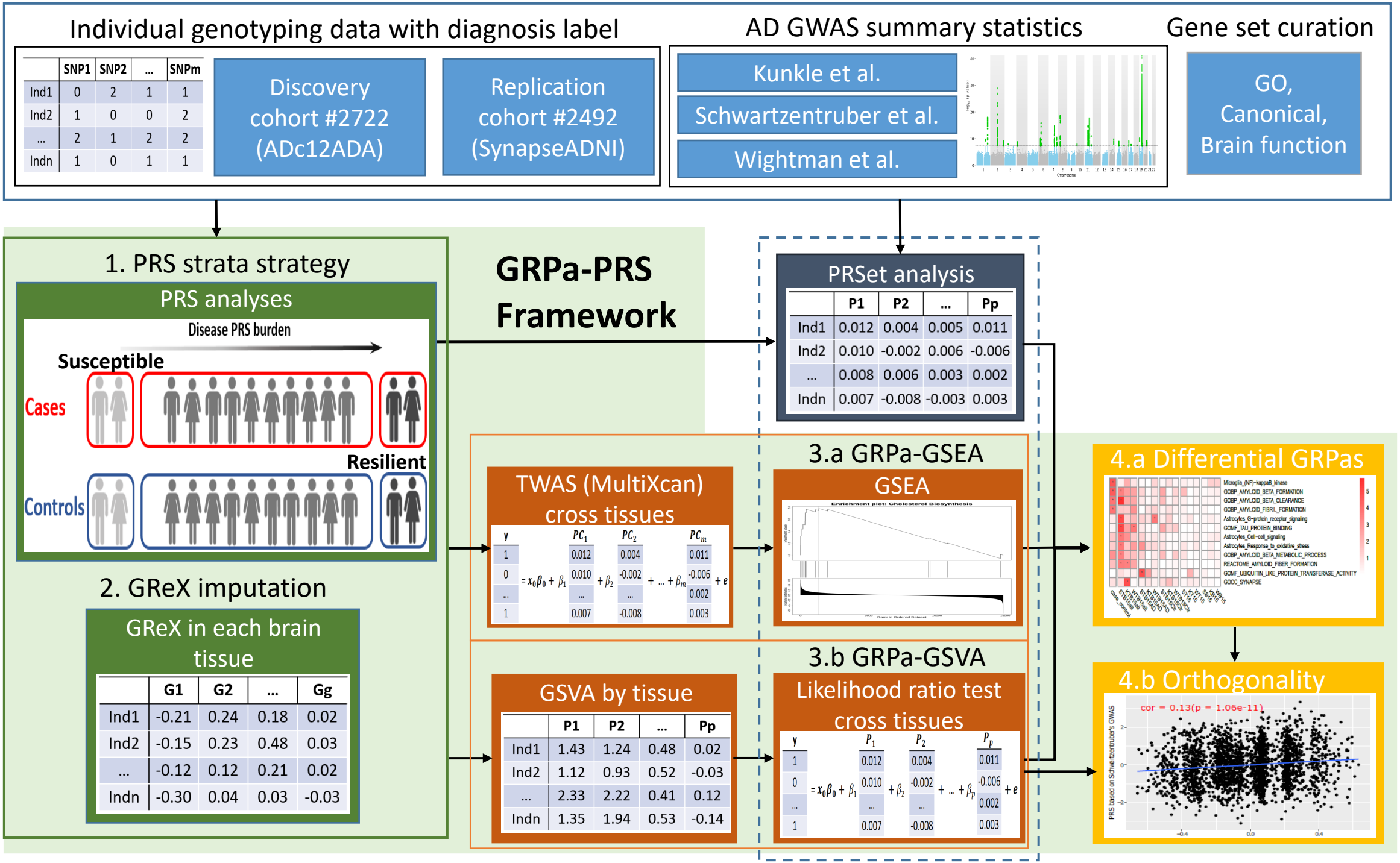
K: stratify by PRS based on Kunkle's GWAS;

S: stratify by PRS based on Schwartzentruber's GWAS;

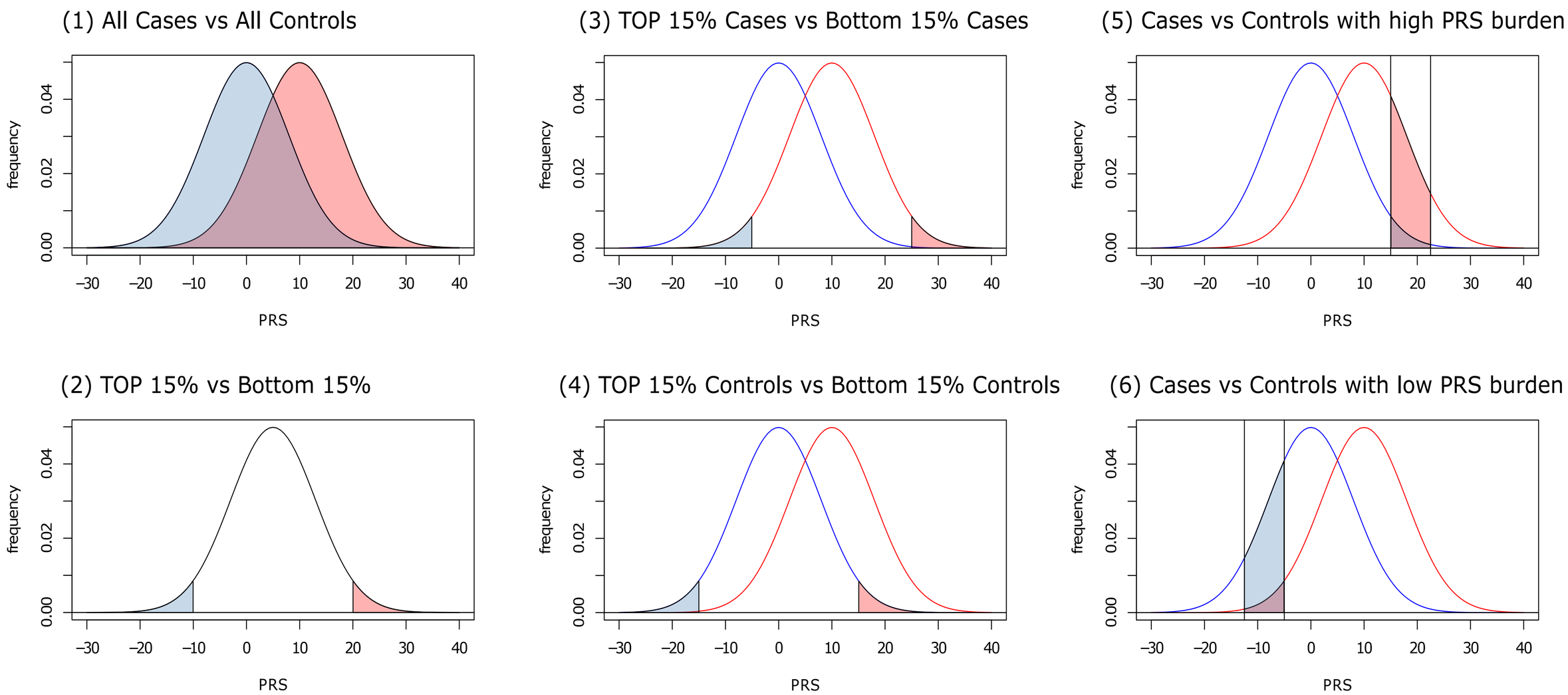
W: stratify by PRS based on Wightman's GWAS;

Data source:

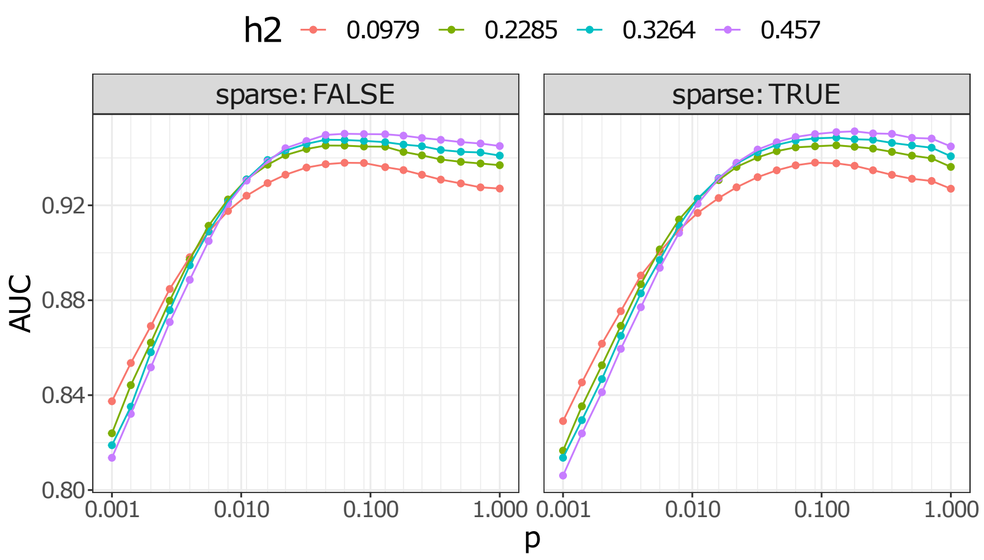
ADc12: National Institute of Aging/Late Onset Alzheimer's Disease Study (NIA/LOAD) cohort consents 1 and 2; GenADA: Multi-Site Collaborative Study for Genotype-Phenotype Associations in Alzheimer's Disease; ROS/MAP: Religious Orders Study and Memory and Aging Project; Mayo: MayoRNAseq; MSBB: Mount Sinai Brain Bank;


Individual genotyping data with diagnosis label
SNP1 SNP2 ... SNPm
Ind1 0 2 1 1
Ind2 1 0 0 2
... 2 1 2 2
Indn 1 0 1 1
Discovery cohort #2722 (ADc12ADA)
Replication cohort #2492 (SynapseADNI)
AD GWAS summary statistics
Kunkle et al.
Schwartzentruber et al.
Wightman et al.
Gene set curation
GO, Canonical, Brain function
GRPa-PRS Framework
1. PRS strata strategy
PRS analyses
Disease PRS burden
Susceptible
Cases
Resilient
Controls
2. GReX imputation
GReX in each brain tissue
G1 G2 ... Gg
Ind1 -0.21 0.24 0.18 0.02
Ind2 -0.15 0.23 0.48 0.03
... -0.12 0.12 0.21 0.02
Indn -0.30 0.04 0.03 -0.03
PRSet analysis
P1 P2 ... Pp
Ind1 0.012 0.004 0.005 0.011
Ind2 0.010 -0.002 0.006 -0.006
... 0.008 0.006 0.003 0.002
Indn 0.007 -0.008 -0.003 0.003
TWAS (MultiXcan) cross tissues
y PC1 PC2 PCm
1 0.012 0.004 0.011
0 0.010 -0.002 -0.006
... ... 0.002
1 0.007 -0.008 0.003
= x0β0 + β1 + β2 + ... + βm + e
3.a GRPa-GSEA
GSEA
Enrichment plot: Cholesterol Biosynthesis
GSVA by tissue
P1 P2 ... Pp
Ind1 1.43 1.24 0.48 0.02
Ind2 1.12 0.93 0.52 -0.03
... 2.33 2.22 0.41 0.12
Indn 1.35 1.94 0.53 -0.14
3.b GRPa-GSVA
Likelihood ratio test cross tissues
y P1 P2 Pp
1 0.012 0.004 0.011
0 0.010 -0.002 -0.006
... ... 0.002
1 0.007 -0.008 0.003
= x0β0 + β1 + β2 + ... + βp + e
4.a Differential GRPas
Microglia_(NF)-kappaB_kinase
GOBP_AMYLOID_BETA_FORMATION
GOBP_AMYLOID_BETA_CLEARANCE
GOBP_AMYLOID_FIBRIL_FORMATION
Astrocytes_G-protein_receptor_signaling
GOMF_TAU_PROTEIN_BINDING
Astrocytes_Cell-cell_signaling
Astrocytes_Response_to_oxidative_stress
GOBP_AMYLOID_BETA_METABOLIC_PROCESS
REACTOME_AMYLOID_FIBER_FORMATION
GOMF_UBIQUITIN_LIKE_PROTEIN_TRANSFERASE_ACTIVITY
GOCC_SYNAPSE
4.b Orthogonality
cor = 0.13(p = 1.06e-11)
PRS based on Schwartzentruber's GWAS

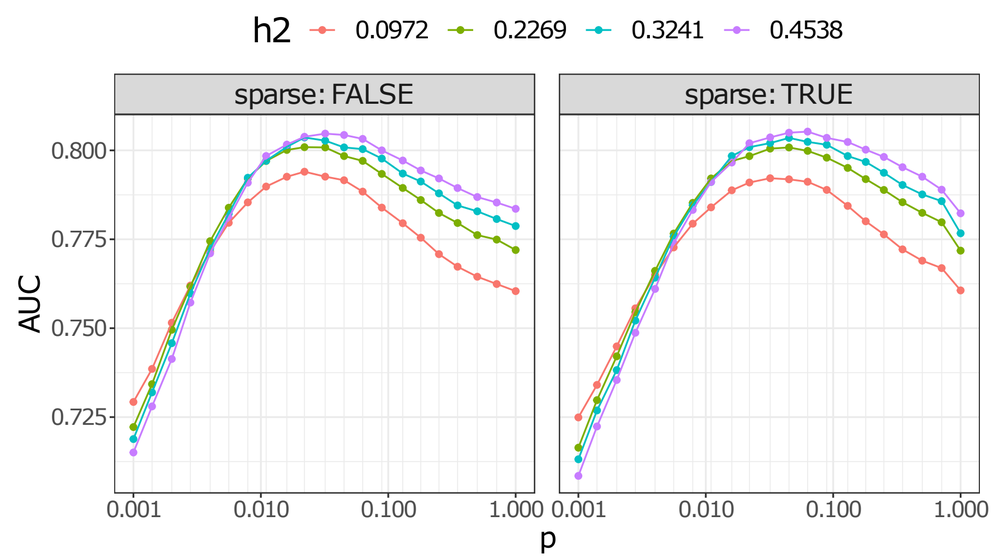
(A) Subgroup design



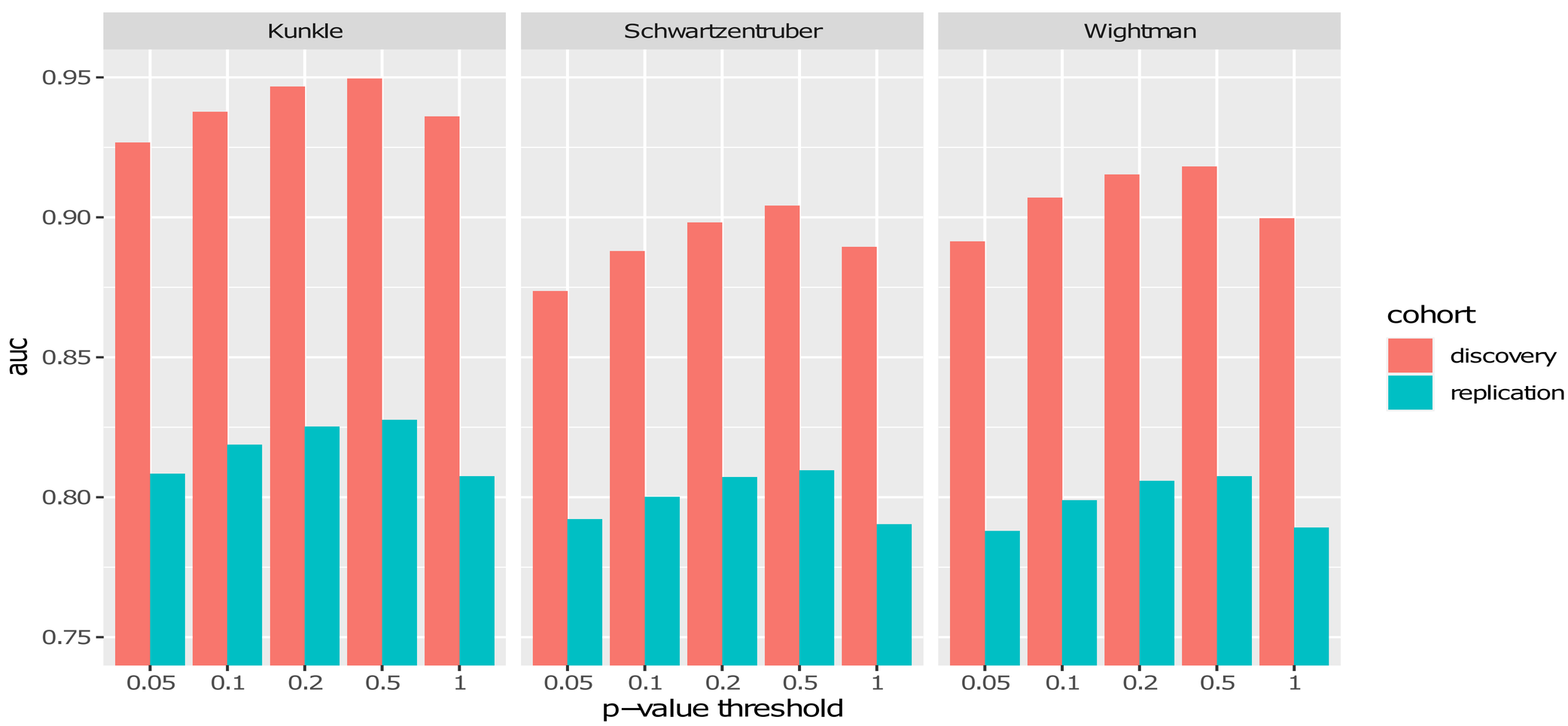
(B) Hyperparameter choice of discovery cohort



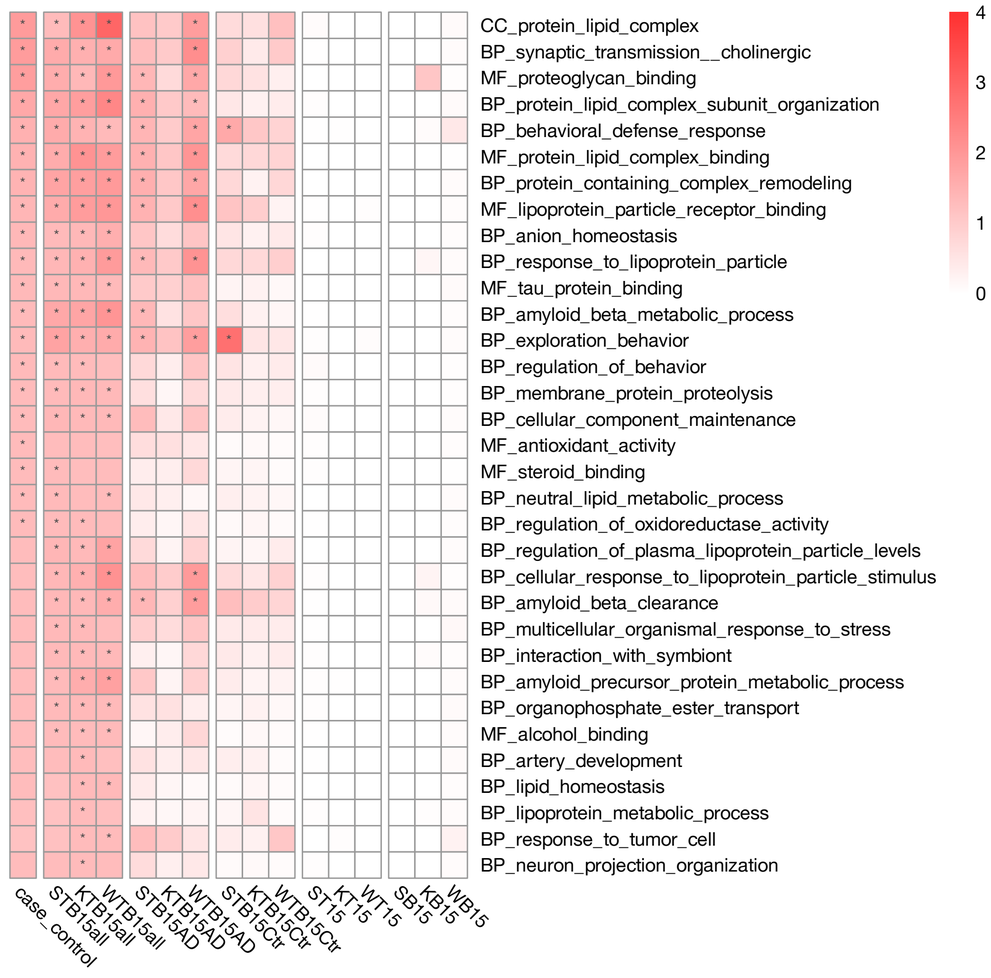
(C) Hyperparameter choice of replication cohort



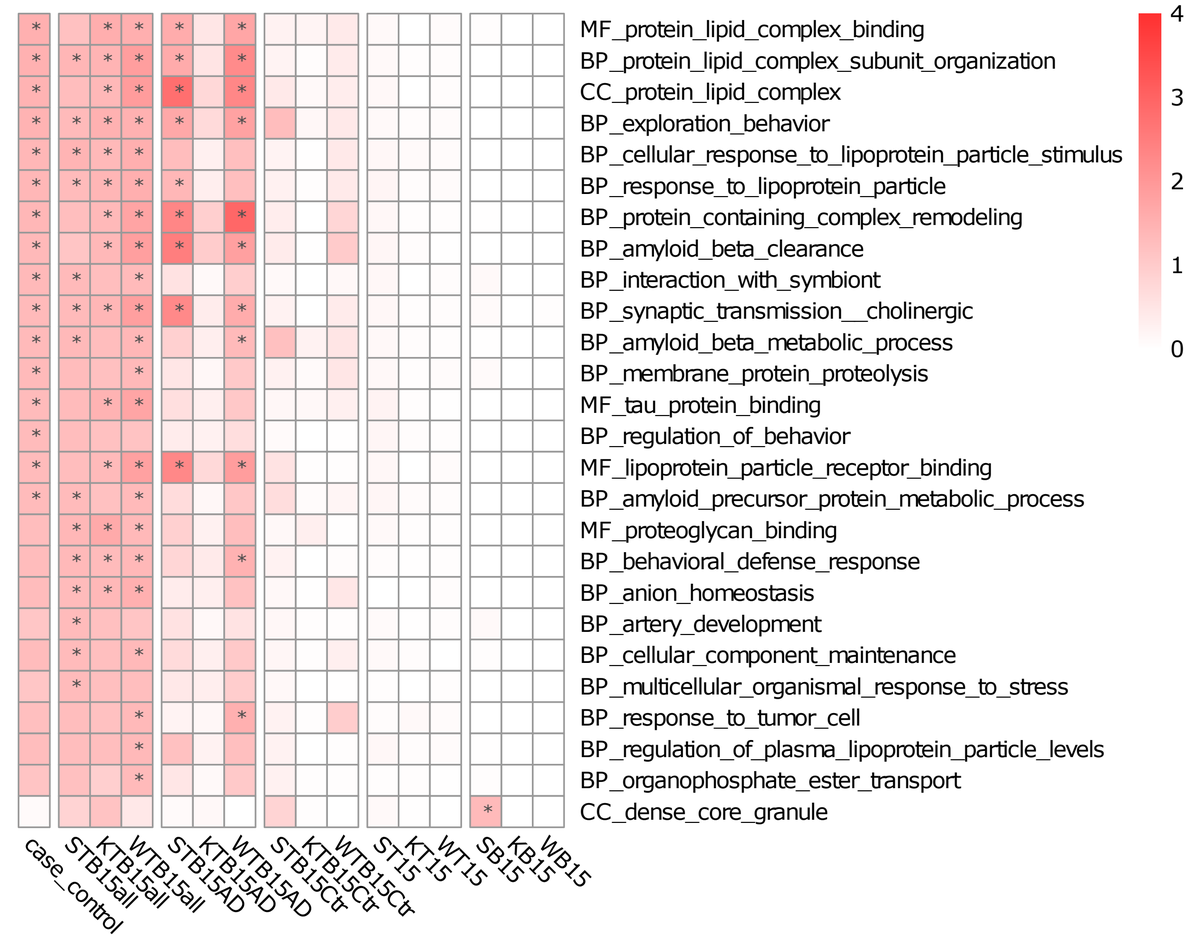
(D) PRS prediction performance vs P-value threshold



(A) GO GRPa identified in discovery dataset Model 1



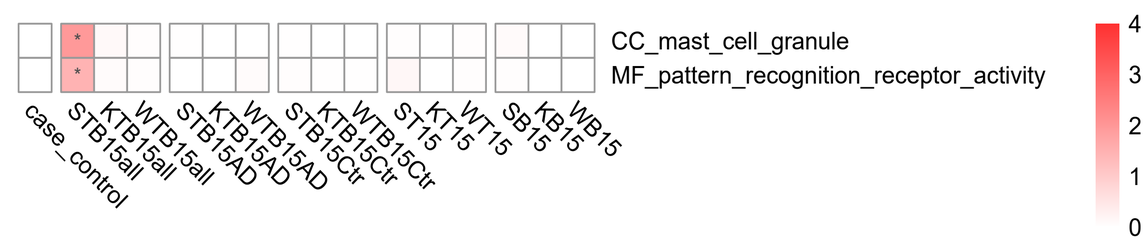
(B) GO GRPa identified in replication dataset Model 1



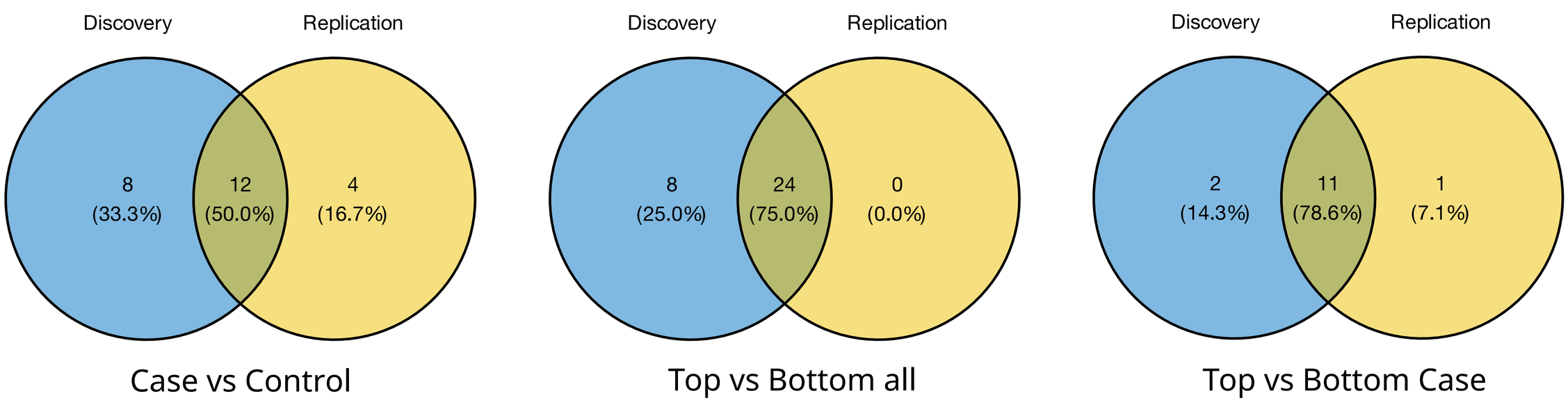
(C) GO GRPa identified in discovery dataset Model 2

No significant GRPa identified in this condition.

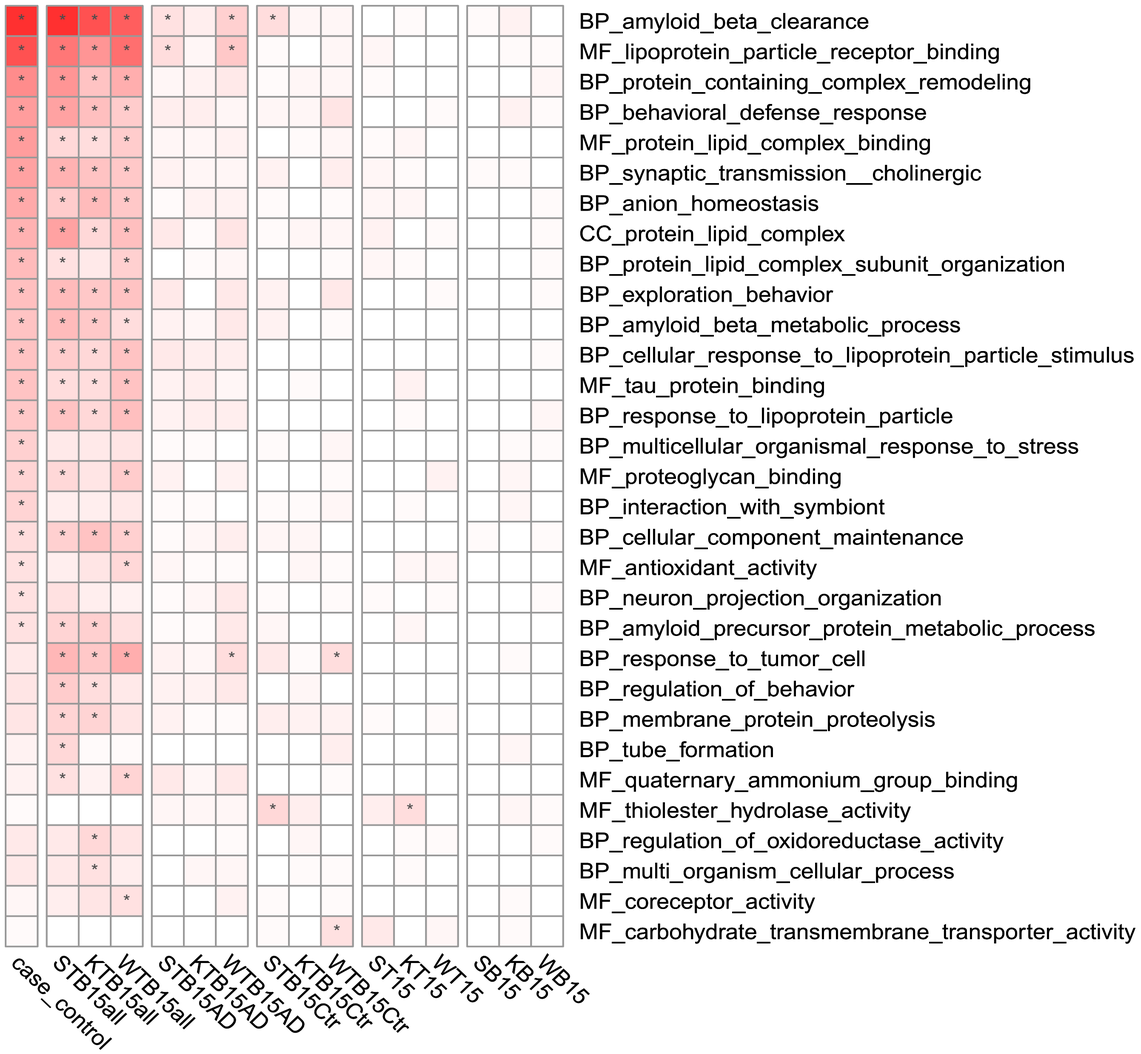
(D) GO GRPa identified in replication dataset Model 2



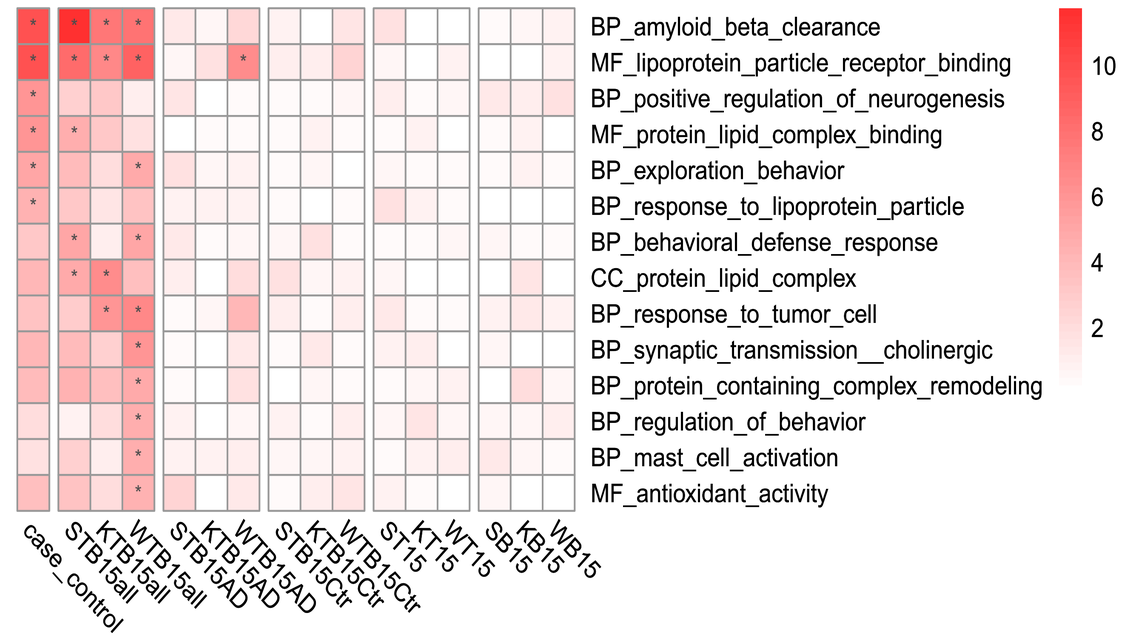
(E) Venn plot for overlap signals between discovery and replication datasets in Model 1



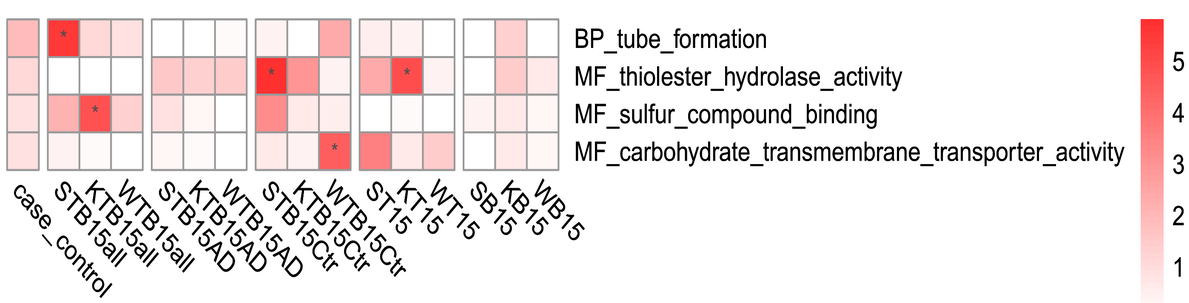
(A) GO GRPa identified in discovery dataset Model 1



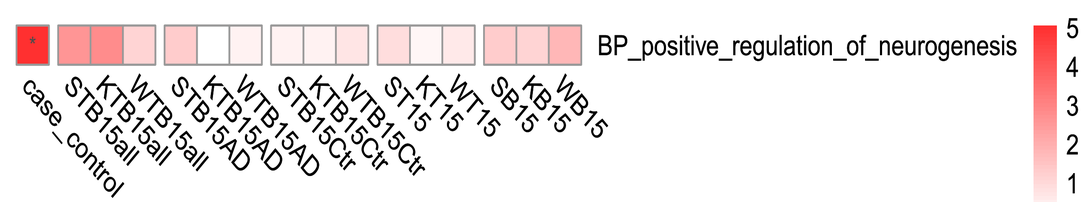
(B) GO GRPa identified in replication dataset Model 1



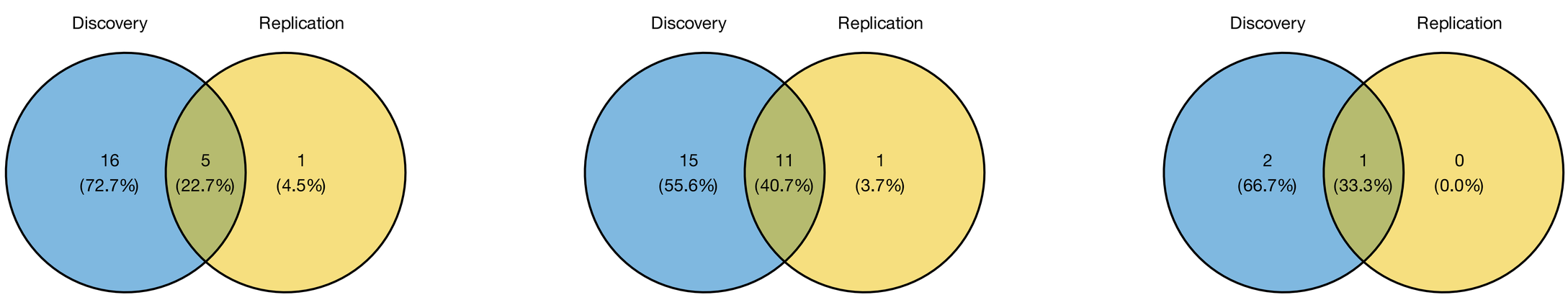
(C) GO GRPa identified in discovery dataset Model 2



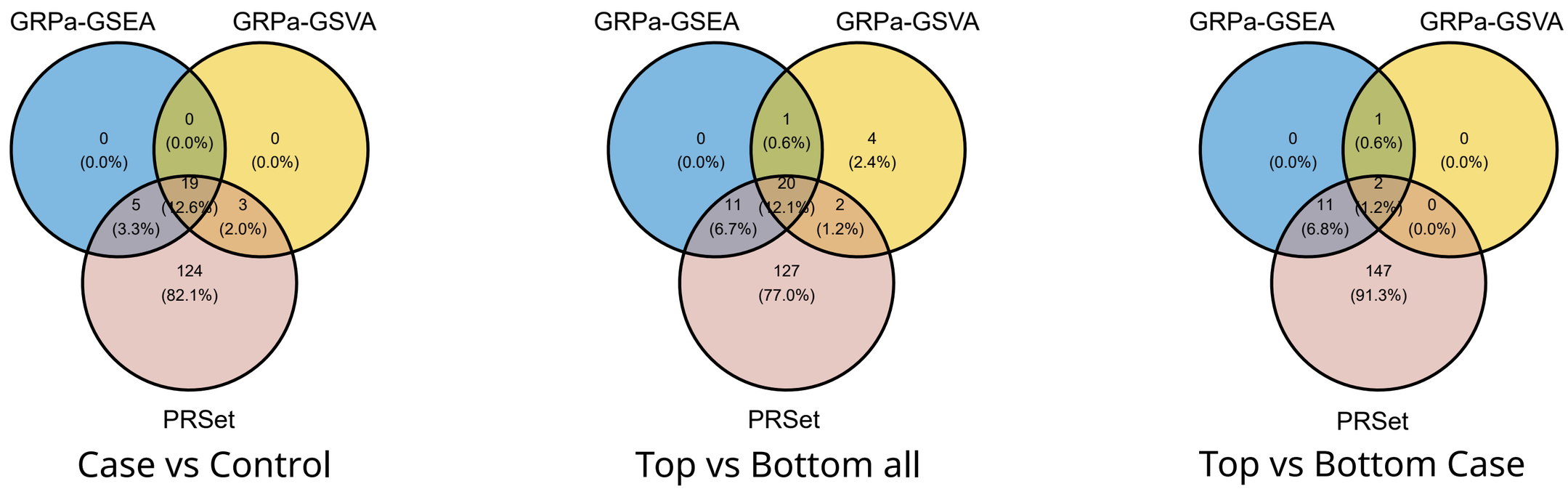
(D) GO GRPa identified in replication dataset Model 2



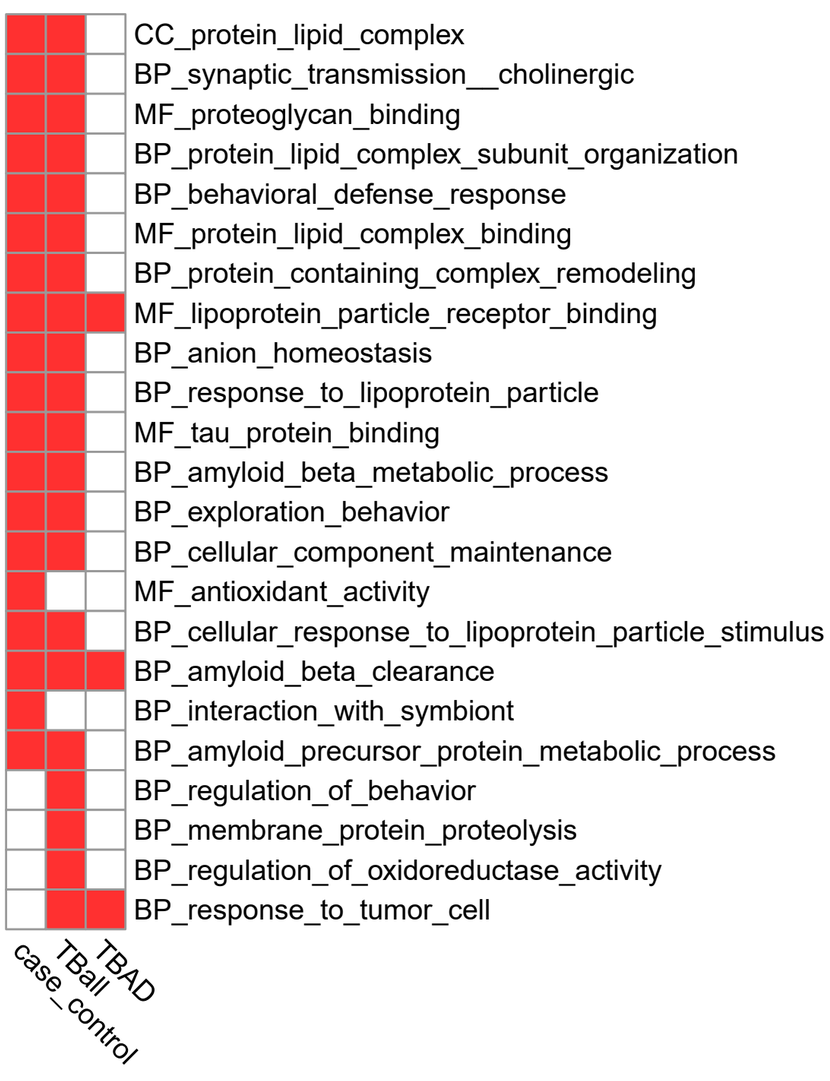
(E) Venn plot for overlapping signals between discovery and replication datasets in Model 1



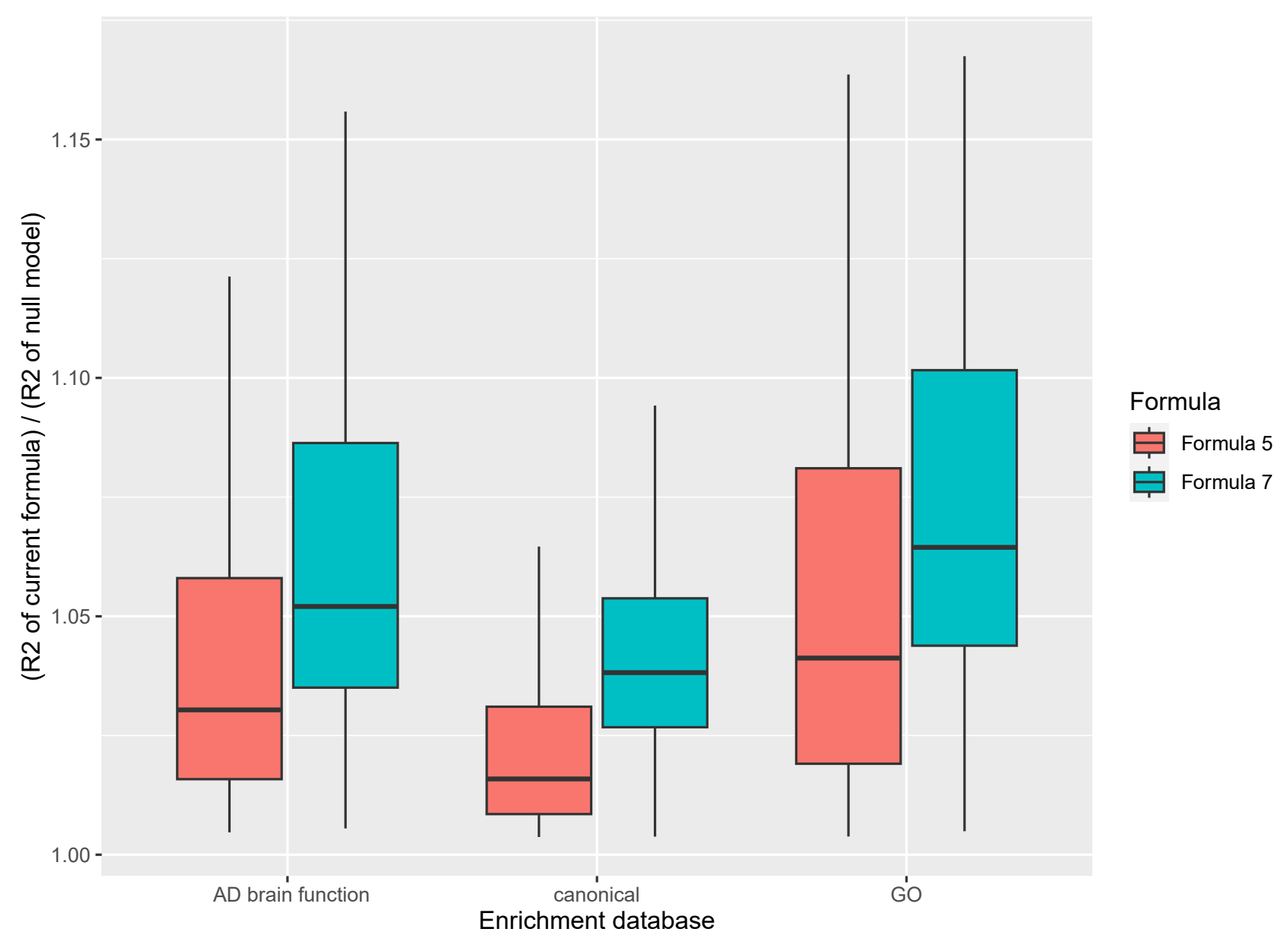
(A) Venn plot for overlapping signals between different approaches



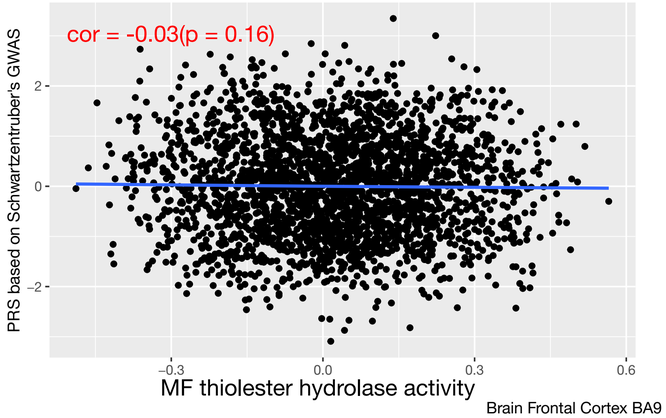
(B) Overlapping pathways between GRPa-GSEA and GRPa-GSVA



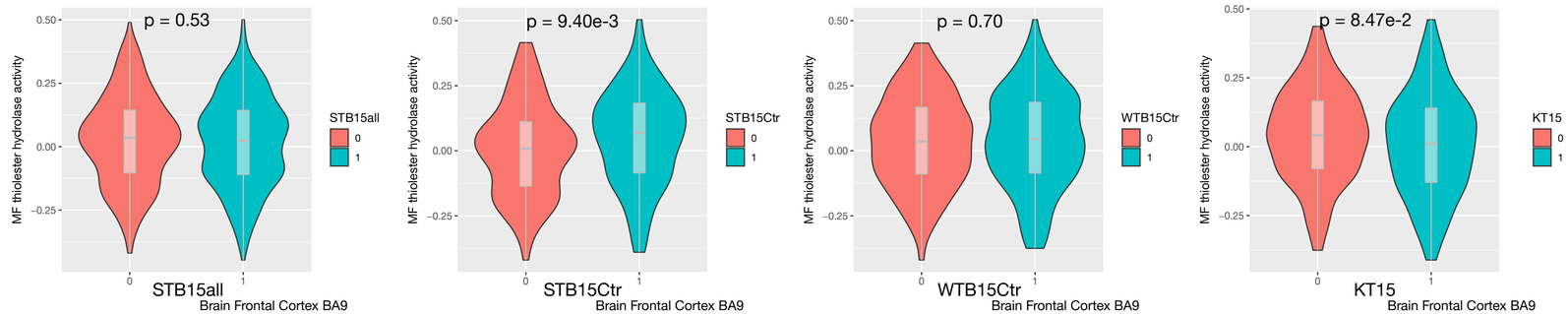
(C) R2 improvement ratio among different enrichemnt database



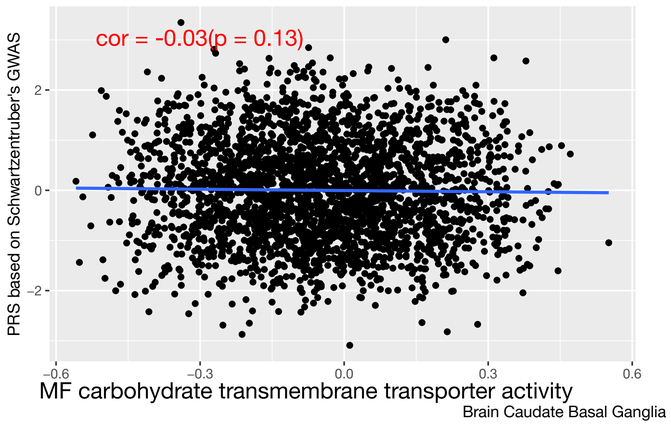
(A) Correlation between thiolester hydrolase activity and sPRS



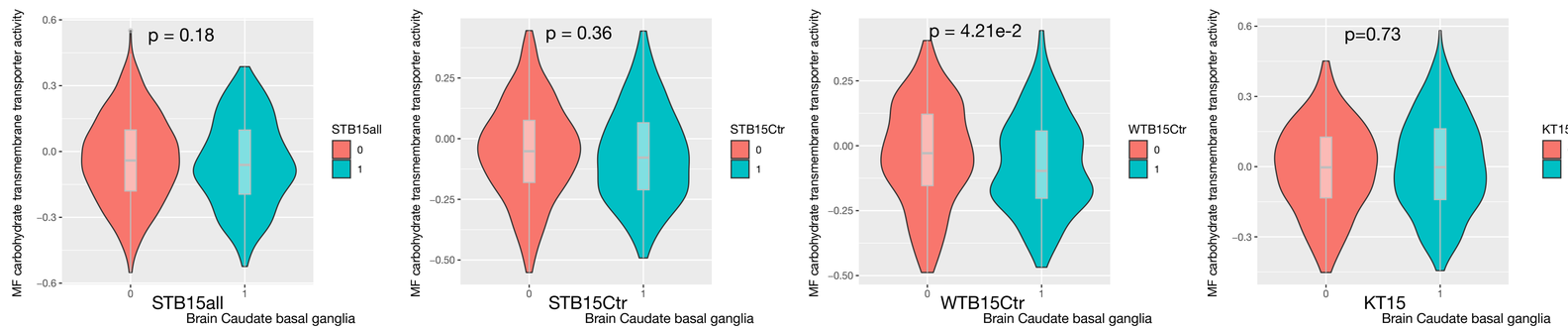
(B) Violin plots for thiolester hydrolase activity in subgroup comparison



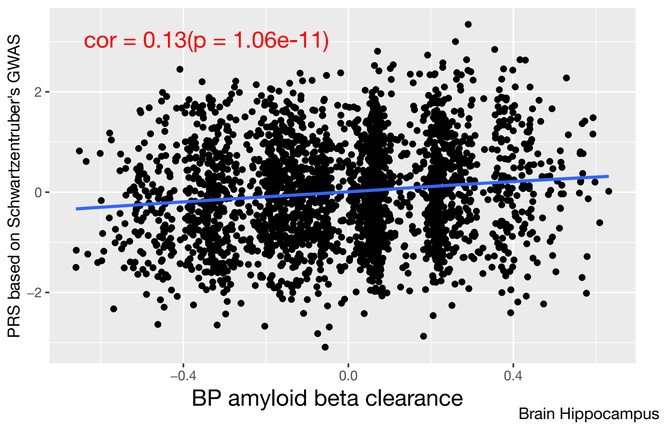
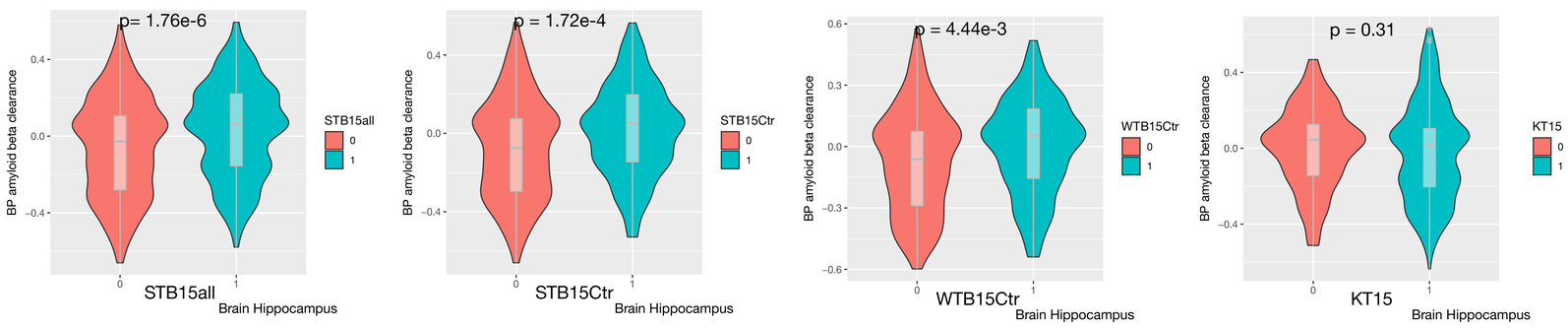
(C) Correlation between carbohydrate transmembrane transporter activity and sPRS



(D) Violin plots for carbohydrate transmembrane transporter activity in subgroup comparison



(E) Correlation between amyloid beta clearance and sPRS



(F) Violin plots for amyloid beta clearance in subgroup comparison