Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Causal association and molecular mechanisms linking insomnia to late- onset Alzheimer's disease: an integrated Mendelian randomization and transcriptomic analysis

Yuanyuan Jiang
Affiliated Hospital of Nanjing University of Chinese Medicine
Chengyong Liu
Affiliated Hospital of Nanjing University of Chinese Medicine
Lu Lu
Liyang Hospital of Chinese Medicine
Xiaoni Zhang
Nanjing University of Chinese Medicine
Shan Qin
Affiliated Hospital of Nanjing University of Chinese Medicine
Xiaomeng Zhao
The Affiliated Suzhou Hospital of Nanjing University Medical School
Wenzhong Wu
fsyy00664@njucm.edu.cn

Affiliated Hospital of Nanjing University of Chinese Medicine

Research Article

Keywords: insomnia, Alzheimer's disease, Mendelian randomization, transcriptomics, weighted gene co-expression network analysis, machine learning

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11643513/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Causal association and molecular mechanisms linking insomnia to late-onset Alzheimer's disease: an integrated Mendelian randomization and transcriptomic analysis

Yuanyuan Jiang[1,2], Chengyong Liu[1], Lu Lu[4], Xiaoni Zhang[3], Shan Qin[1], Xiaomeng Zhao[5], Wenzhong Wu[1*]

[1] Department of Acupuncture and Rehabilitation, Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China

[2] Jiangsu Province Hospital with Integration of Chinese and Western Medicine, Nanjing, China

[3] Nanjing University of Chinese Medicine, Nanjing, China

[4] Liyang Hospital of Chinese Medicine, Changzhou, China

[5] The Affiliated Suzhou Hospital of Nanjing University Medical School, Suzhou, China

*Corresponding author: Wenzhong Wu - fsyy00664@njucm.edu.cn

Abstract

Background: Insomnia is consistently associated with Alzheimer's disease (AD) in observational studies, but the causal relationship and the underlying molecular mechanisms remain unclear. This study aimed to investigate the causal relationship between insomnia and late-onset AD and to identify molecular pathways and candidate genes that may contribute to this association by integrating genetic causal inference with transcriptomics.

Methods: Two-sample Mendelian randomization (MR) was performed using 28 single-nucleotide polymorphisms as instruments, with insomnia genome-wide association study summary statistics from UK Biobank (336,965 participants of European ancestry) and late-onset AD data from FinnGen (11,359 cases and 508,613 controls). Inverse-variance weighting was the primary method. Robustness was assessed using MR-Egger, weighted median, simple mode, and weighted mode estimates, together with pleiotropy, heterogeneity, and leave-one-out analyses. GEO datasets GSE122063 and GSE110226 were integrated for differential expression, weighted gene co-expression network, protein interaction, and functional enrichment analyses. Lasso regression and support vector machine recursive feature elimination were used to identify key candidate genes, whose discrimination was assessed using receiver operating characteristic curves.

Results: Inverse-variance weighting associated genetically predicted insomnia with higher odds of late-onset AD (OR = 3.36, 95% CI 1.02–10.99; P = 0.045). Other MR methods yielded directionally consistent but imprecise estimates. The MR-Egger intercept test found no evidence of directional pleiotropy (P = 0.939), and Cochran's Q test was not statistically significant (P = 0.062). No individual instrument dominated the leave-one-out results. Transcriptomic analysis identified 1,616 differentially expressed genes and eight co-expression modules. MEpink, MEblack, and MEbrown showed the strongest associations with AD. Their intersection with 2,330 insomnia-related genes yielded 35 candidates, enriched mainly in neuroinflammation, cell adhesion, and lipid metabolism. Both machine learning algorithms selected seven key candidates. All seven were differentially expressed between AD and controls in the combined dataset, with individual AUCs above 0.75.

Conclusion: Genetic liability to insomnia may be associated with an increased risk of late-onset AD. Neuroinflammation and lipid metabolism may link the two conditions, and seven candidate genes warrant further study. Given the borderline MR result, wide confidence interval, tissue heterogeneity, and absence of independent external validation, these findings require confirmation in larger multi-ancestry cohorts, independent transcriptomic datasets, and experimental studies.

Keywords: insomnia; Alzheimer's disease; Mendelian randomization; transcriptomics; weighted gene co-expression network analysis; machine learning

Background

Observational studies and meta-analyses associate insomnia and poor sleep quality with cognitive decline, all-cause dementia, and Alzheimer's disease (AD) [1,2]. Establishing insomnia as a modifiable cause of AD requires more than this association. Sleep disturbances are common in older adults, and AD, one of the most common neurodegenerative diseases, can disrupt the sleep–wake cycle. AD pathology also develops years before clinical symptoms. Depression, vascular risk factors, medication use, and socioeconomic circumstances further complicate the relationship because they may affect both sleep and cognition.

Sleep architecture and continuity are abnormal in AD patients, according to a systematic review and meta-analysis of polysomnographic studies [5]. Reviews of the underlying biology describe a potentially bidirectional relationship: glymphatic clearance, amyloid-β and tau homeostasis, circadian rhythms, and neuroimmune activity all provide possible connections [3,4]. Blood–brain barrier and choroid plexus dysfunction and altered lipid metabolism may also participate in the interaction between disturbed sleep and neurodegeneration. The relationship is biologically plausible, but its causal direction and long-term effect are still unresolved.

Previous two-sample Mendelian randomization (MR) studies generally found no clear causal effect of insomnia or other sleep phenotypes on AD risk [6], leaving inconsistent evidence. MR uses germline genetic variants as instruments; under specified assumptions, it reduces the influence of confounding and reverse causation. Molecular mechanisms remain outside what a genetic association alone can establish. For that question, brain transcriptomics and network analyses allow examination of co-expression modules, protein interactions, and functional pathways. These analyses, together with machine learning, can supply testable hypotheses to complement MR evidence.

Our analysis brought these sources of evidence together for insomnia and late-onset AD. We used two-sample MR to examine whether genetic liability to insomnia was associated with AD risk. From two public AD brain transcriptomic datasets, we sought shared modules and pathways and candidate genes worth validating. The analyses comprised weighted gene co-expression networks, protein interaction networks, functional enrichment, and two complementary machine learning algorithms. We organized the report according to STROBE-MR [7].

Methods

Study design

This study used a three-stage integrated design to examine the relationship between insomnia and AD, from genetic causal inference and molecular mechanism exploration to therapeutic target identification. First, bidirectional two-sample MR used genetic variants

as instruments to assess causality while addressing confounding and reverse causation inherent in observational research. Second, AD transcriptomic data from the Gene Expression Omnibus (GEO) were integrated to identify insomnia-related co-expression modules using weighted gene co-expression network analysis (WGCNA), followed by construction of a protein–protein interaction (PPI) network. Third, Lasso regression and support vector machine recursive feature elimination (SVM-RFE) were used to cross-validate key target selection and assess diagnostic performance. This design followed a translational sequence of genetic association, mechanistic investigation, and target validation, with reporting in accordance with STROBE-MR [7]. Only publicly available, de-identified summary-level and transcriptomic data were analyzed; no human participants were recruited, and therefore no ethical approval or informed consent to participate was required (see Declarations).

Mendelian randomization

Data sources

Exposure data were obtained from UK Biobank genome-wide association study (GWAS) summary statistics for insomnia (identifier ukb-a-13), comprising 336,965 participants of European ancestry. This dataset used the self-reported phenotype "sleeplessness/insomnia" and underwent standardized quality control [8].

Outcome data were obtained from FinnGen release R13 for late-onset Alzheimer's disease (identifier finn-b-AD_LO), including 11,359 cases and 508,613 controls from Finland (519,972 participants in total; 11,359 cases after genotype quality control). Cases were identified through hospital discharge and cause-of-death registers using International Classification of Diseases, Tenth Revision (ICD-10) codes F00.1 (dementia in late-onset AD), F00.10 (dementia in late-onset AD without additional behavioral disturbance), F00.10*G30.1, G30.1 (late-onset AD), and corresponding combinations (G30.1+F00.10). Controls were Finnish participants without an AD diagnosis who were retained after genotype quality control and exclusion of related samples.

We selected SNPs associated with insomnia at $P < 5 \times 10^{-8}$. Linkage disequilibrium clumping used the European 1000 Genomes reference panel, with $r^2 < 0.001$ and a 10,000-kb window. F statistics were calculated for instrument strength, and SNPs with $F < 10$ were excluded to reduce weak-instrument bias. These steps left 28 independent SNPs for the MR analysis.

MR methods

Five complementary MR methods were used. The primary inverse-variance weighted (IVW) analysis combined SNP-specific Wald ratios by meta-analysis, assuming that all variants satisfied the core instrumental variable assumptions of relevance, independence, and exclusion restriction, to provide the most statistically efficient causal estimate. Sensitivity analyses comprised MR-Egger regression, which allows directional pleiotropy

and uses the intercept to test average pleiotropic effects and the slope to estimate the causal effect; the weighted median method, which assumes that at least 50% of genetic variants are valid instruments and is robust to outliers; and simple and weighted mode estimators, which assume that the largest cluster of variants contains valid instruments. Leave-one-out analyses repeated MR after removing each SNP in turn to assess whether particular variants drove the overall estimate. Funnel plots were used to inspect the symmetry of SNP effect estimates as an additional assessment of possible pleiotropy [9–11].

Pleiotropy and heterogeneity

We assessed horizontal pleiotropy with the MR-Egger intercept test, which examines effects of genetic variants on the outcome through pathways independent of the exposure. An intercept differing from zero at $P < 0.05$ was taken as evidence of directional pleiotropy. Cochran's Q tested heterogeneity across SNP-specific causal estimates. A Q-test P value < 0.05 indicated heterogeneity, potentially arising from pleiotropy or violations of the instrumental variable assumptions. MR analyses were run in R with TwoSampleMR version 0.6.4 [9,12].

Transcriptomic data integration

Data acquisition

AD-related transcriptomic datasets were retrieved from GEO. Eligibility criteria were Homo sapiens, brain tissue, inclusion of AD cases and normal controls, and a sample size of at least 10. The sample-size threshold was lowered to accommodate the availability of rare-disease datasets and ensure data diversity.

Two independent datasets were included. GSE122063 used the Agilent-039494 SurePrint G3 Human GE v2 8x60K Microarray platform (GPL16699) and comprised 44 normal control and 56 AD samples, for a total of 100. The tissues were prefrontal and temporal cortex. Although the original study also included vascular dementia, only AD and control samples were extracted here. GSE110226 used the Rosetta/Merck Human RSTA Custom Affymetrix 2.0 microarray platform (GPL10379), comprising six normal controls and seven AD samples, for a total of 13. Tissue was obtained from the lateral ventricular choroid plexus. The original study also included frontotemporal dementia (four samples) and Huntington's disease (three samples); only AD and control samples were used in this analysis [13,14].

Raw CEL files or expression matrices were downloaded using GEOquery in R. Platform-specific preprocessing was then applied. GSE122063 underwent standard Agilent background correction, Lowess normalization, and probe annotation, whereas GSE110226 underwent Affymetrix background correction, robust multi-array average (RMA) normalization, and probe annotation. Probes were mapped to official gene symbols; where multiple probes mapped to the same gene, the probe with the greatest

variability was retained. ComBat in the sva package was then used to remove technical differences between datasets while retaining biological variation. Principal component analysis (PCA) and sample clustering were used to visualize the effectiveness of batch correction [15].

Differential expression

Differential expression was assessed in the batch-corrected combined dataset. Linear models fitted with limma in R estimated expression changes in AD relative to controls [16]. Differentially expressed genes were defined by $|\log_2(\text{fold change})| > 0.58$ and a Benjamini–Hochberg-adjusted P value (false discovery rate, FDR) < 0.05 [17]. A heatmap of the 50 most significant genes, ranked by adjusted P value, was generated with pheatmap to display sample clustering and expression patterns. Volcano plots generated with ggplot2 identified significantly upregulated and downregulated genes.

Weighted gene co-expression network analysis

WGCNA was used to identify gene co-expression modules associated with AD status and insomnia. Hierarchical clustering and dynamic tree cutting partitioned the gene dendrogram into modules. For each module, the first principal component was defined as the module eigengene (ME), representing its overall expression pattern. Pearson correlations and corresponding P values between module eigengenes and AD phenotype identified modules associated with disease status [18].

Insomnia-related genes were retrieved from the GeneCards database (https://www.genecards.org/) using “insomnia” or “sleep initiation and maintenance disorders,” with a relevance score > 5. These genes were intersected with members of significant WGCNA modules to obtain candidate genes linking insomnia and AD. Modules were ranked by the significance of their associations with the phenotype ($P < 0.05$), and the three most strongly correlated modules were selected for further analysis.

Protein–protein interaction analysis

A PPI network was constructed to examine interactions among proteins encoded by insomnia–AD-associated genes. Genes shared by the key WGCNA modules and the insomnia-related set were submitted to STRING version 12.0 (https://string-db.org/). The organism was restricted to Homo sapiens, the interaction confidence threshold was > 0.7, and disconnected nodes were hidden [19].

Interaction data were downloaded in TSV format and imported into Cytoscape version 3.9.1 for visualization. Densely connected subnetworks were identified with the MCODE plugin using a degree cutoff of 5 [20].

Functional enrichment

Functional annotation and pathway enrichment were performed on the shared insomnia–AD genes to explore potential biological mechanisms. Gene Ontology (GO) enrichment

used hypergeometric tests in clusterProfiler version 4.18.4 for biological processes, molecular functions, and cellular components. All protein-coding genes were used as the background. Benjamini–Hochberg correction was applied, with FDR < 0.05 defining significance. KEGG enrichment was also performed with clusterProfiler, focusing on neurodegenerative disease, immune and inflammatory, and metabolic pathways, using the same significance threshold. Bar and dot plots generated with ggplot2 displayed significant GO terms and KEGG pathways ranked by gene ratio or enrichment significance [21].

Machine learning selection of key genes

Two complementary machine learning methods were used to select key targets from the intersecting gene set. Lasso regression performed variable selection through L1 regularization. Ten-fold cross-validation in glmnet selected the optimal λ, and genes with nonzero coefficients were retained. SVM-RFE iteratively removed low-weight features using recursive feature elimination in caret. Genes selected by both methods were defined as key targets. Receiver operating characteristic (ROC) curves generated with pROC assessed diagnostic performance; an area under the curve (AUC) > 0.75 was considered indicative of good diagnostic value [22–24].

Results

Causal effect of insomnia on AD risk

Main MR results

MR analysis indicated that genetically predicted insomnia significantly increased the risk of late-onset AD (Figure 1). The primary IVW estimate supported a significant causal association (odds ratio [OR] = 3.36, 95% confidence interval [CI] 1.02–10.99; P = 0.045). This estimate was interpreted as an approximately 236% higher AD risk in individuals genetically predisposed to insomnia than in those without such a predisposition.

Sensitivity estimates were directionally consistent with the primary analysis but had wide confidence intervals. The weighted median estimate was OR = 4.54 (95% CI 0.95–21.70; P = 0.06), approaching statistical significance. The MR-Egger estimate was OR = 2.96 (95% CI 0.10–89.77; P = 0.54); its interval included the null value and was imprecise. The simple mode estimate was OR = 4.66 (95% CI 0.29–74.66; P = 0.29), and the weighted mode estimate was OR = 4.88 (95% CI 0.58–40.81; P = 0.15). Agreement in effect direction across the five methods strengthened the causal interpretation, although some methods did not reach conventional significance because of limited statistical power (Table 1).

Sensitivity and robustness analyses

The MR-Egger intercept was 0.0017 (standard error [SE] = 0.022; P = 0.939). The nonsignificant intercept was interpreted as indicating no directional pleiotropy and supporting instrument validity.

Cochran's Q was 39.09 with 27 degrees of freedom (P = 0.062). This result did not reach conventional significance and was interpreted as indicating no heterogeneity among SNP-specific causal estimates, consistent with the homogeneity assumption of IVW.

In leave-one-out analyses, IVW was repeated after excluding each of the 28 SNPs. All OR estimates ranged from 2.5 to 4.5 and remained above 1.0, indicating that no single SNP drove the overall association (Figure 1B).

Scatter, funnel, and forest plots

The scatter plot displayed SNP effects on insomnia on the x-axis and on AD on the y-axis. All five MR regression slopes were positive, visually indicating a positive causal association (Figure 1A). The funnel plot showed approximately symmetrical SNP estimates around the overall effect, with no obvious signal of publication bias or pleiotropy (Figure 1C). The forest plot summarized the MR-Egger and IVW estimates with their 95% confidence intervals, showing that all point estimates lay above the null value of OR = 1 (Figure 1D).

Transcriptomic characteristics of AD

Differential expression

Integration of GSE122063 and GSE110226 after ComBat batch correction identified 1,616 differentially expressed genes meeting $|\log_2 FC| > 0.58$ and FDR < 0.05. Of these, 961 were downregulated (control > AD) and 655 were upregulated (AD > control).

The 50 differentially expressed genes with the greatest expression variability were plotted in a heatmap. AD and control samples formed distinct clusters, which was interpreted as evidence of successful integration after batch correction (Figure 2A). The volcano plot showed the distribution of genes by $\log_2 FC$ and $-\log_{10}(P)$. Upregulated and downregulated genes were broadly balanced in number, without an extreme imbalance (Figure 2B).

Assessment of batch correction

PCA showed clear separation of the datasets along the first principal component before batch correction and mixing afterward, indicating removal of technical batch effects while retaining biological signals. The sample dendrogram with the clinical trait heatmap of the combined dataset showed that corrected samples clustered by AD status rather than by dataset (Figure 2C).

WGCNA identification of insomnia-related modules

Network topology

WGCNA identified eight co-expression modules. The gene dendrogram and module colors indicated module membership (Figure 2D). The sample dendrogram with a clinical phenotype heatmap identified no outliers requiring exclusion; all 113 samples were retained for network construction (Figure 2C).

Module–trait relationships

MEpink showed the strongest association with AD ($r = -0.64$, $P = 3 \times 10^{-14}$), followed by MEblack ($r = 0.55$, $P = 3 \times 10^{-10}$) and MEbrown ($r = 0.52$, $P = 4 \times 10^{-9}$) (Figure 2E).

GeneCards searches yielded 2,330 insomnia-related genes. Intersection with the three selected modules identified 35 insomnia–AD candidate genes (Figure 2F). The high-confidence STRING network (score > 0.7) contained 29 nodes (Figure 2G). GO analysis showed significant enrichment of neuroinflammation and cell adhesion processes (Figures 2H and 2I). KEGG analysis indicated that insomnia primarily promoted AD through immune-related and lipid-metabolic pathways, including the Influenza A pathway and the Lipid and atherosclerosis pathway (Figure 2J).

Machine learning identification of key candidate genes

The intersection of Lasso and SVM-RFE identified seven key genes (Figure 3A–D). These genes satisfied the selection criteria of both regularized sparse modeling and maximum-margin classification, suggesting robust potential for AD diagnostic prediction. Functional annotation linked the genes mainly to neuroinflammation and lipid metabolic regulation, consistent with the GO and KEGG findings and supporting central roles for these pathways in the promotion of AD by insomnia.

Validation of key genes

All seven key genes were significantly differentially expressed between AD and controls in the combined GEO dataset (Figure 3E). Six were upregulated in AD and one was downregulated. All individual ROC AUCs exceeded 0.75, meeting the prespecified threshold for good diagnostic performance (Figure 3F).

Discussion

Our genetic and transcriptomic analyses examined different aspects of the insomnia–AD association. IVW linked genetic liability to insomnia to higher late-onset AD risk. The other MR estimates had the same direction, and the sensitivity analyses found no clear evidence of directional pleiotropy or an effect driven by one variant. The overlap between AD co-expression modules and insomnia-related genes implicated neuroinflammation, cell adhesion, and lipid metabolism. Lasso and SVM-RFE both selected seven genes that differed in expression between AD and controls and discriminated between these groups

in the combined dataset. These results identify candidate mechanisms for investigation; they do not establish a causal chain or a clinical diagnostic test.

Interpretation of the MR findings

Our MR estimates had the same direction as the associations reported in observational studies of insomnia and AD [1,2]. They are also compatible with experimental evidence that sleep contributes to metabolic waste clearance and amyloid-β homeostasis [3,4]. Earlier large two-sample MR studies, however, generally found no effect or reached inconclusive results for several sleep phenotypes [6]. Differences in AD datasets and case definitions, exposure phenotypes, instrument numbers, statistical power, or sample overlap could explain the discrepancy. Using registry-defined late-onset AD may reduce some clinical heterogeneity. Self-reported insomnia remains a limitation because it does not clearly distinguish short-term sleep difficulties from chronic insomnia disorder or other sleep conditions.

The IVW confidence interval had a lower bound of 1.02, with a P value close to 0.05. Both features leave considerable uncertainty about the association. Weighted median and mode estimates pointed in the same direction but did not reach conventional significance, and MR-Egger was especially imprecise. We therefore interpret the results as suggesting a possible causal association. They do not confirm that insomnia causes AD. The nonsignificant MR-Egger intercept and Cochran's Q tests also cannot rule out pleiotropy or heterogeneity, as their power may be limited. Both GWAS datasets predominantly included people of European ancestry, restricting generalizability. Other concerns include misclassification in registry diagnoses and the potential effects of population stratification, selection bias, and residual weak-instrument bias.

Neuroinflammation as a potential molecular link

The enrichment of neuroinflammatory and cell adhesion processes points to biology that insomnia and AD may share. Chronic sleep disturbance could increase peripheral and central immune signaling, change microglial responses, and disrupt blood–brain barrier function, making the brain more vulnerable to degeneration. Neuroinflammation forms part of AD pathology rather than simply accompanying its final stages. Innate immune responses interact with amyloid-β and tau pathology, with protective or harmful effects depending on disease stage [25]. The original GSE110226 study found changes in immune responses, cytokine signaling, and cell adhesion in the AD choroid plexus [13]. Those observations agree with the direction of our enrichment results.

The expression data compared AD tissue with control tissue; they did not compare people with and without insomnia. Thus, the enriched pathways describe the overlap of an insomnia-related gene set with AD expression networks. They cannot establish pathway activation in the brains of people with insomnia or mediation of the MR association. Testing these links requires brain or peripheral samples accompanied by objective sleep

measurements. Mediation MR, colocalization, single-cell transcriptomics, and functional experiments could then help resolve the proposed sequence of molecular events.

Lipid metabolism and candidate pathways

Lipid metabolism and vascular processes are possible links between insomnia and AD, as suggested by enrichment of the Lipid and atherosclerosis pathway. Lipid homeostasis supports neuronal membranes, synaptic function, glial responses, and amyloid-β processing, and its disruption is closely associated with AD pathology [26]. Sleep loss also interacts with vascular and metabolic risk factors. These observations suggest a model for testing: genetic and physiological changes associated with long-term insomnia might increase brain vulnerability through interactions among inflammation, lipid metabolism, and vascular function. Enrichment of the Influenza A pathway has a different interpretation. It more likely reflects shared innate immune and antiviral response genes than a role for influenza infection in mediating AD risk in this study. The pathway name establishes neither exposure to a pathogen nor the direction of causality.

Candidate genes and translational implications

Lasso and SVM-RFE select variables using different approaches to dimensionality reduction and feature ranking. Taking their intersection may reduce dependence on either algorithm. All seven selected genes were differentially expressed in the combined dataset, and each had an AUC above 0.75, warranting further biomarker studies. However, feature selection and ROC assessment used the same data. Information leakage and overfitting may therefore have inflated performance. These AUCs do not show that the genes can be measured as clinically useful brain tissue biomarkers or that they are valid therapeutic targets. Evaluation should use independent cohorts stratified by tissue, with features and thresholds fixed in advance. It should assess calibration, sensitivity, specificity, and added predictive value, as well as whether the markers are detectable in cerebrospinal fluid or blood.

An MR estimate cannot tell us how much AD risk would fall if insomnia were treated, even if a causal effect is confirmed. It represents an average effect associated with lifelong genetic liability, which differs from a short course of medication or a behavioral intervention. To establish whether better sleep changes disease progression, intervention trials need long follow-up, cognitive outcomes, and fluid or imaging biomarkers of AD.

Strengths and limitations

The study combined complementary MR estimators, public brain transcriptomic datasets, co-expression networks, and two machine learning methods. Pleiotropy, heterogeneity, and leave-one-out results were reported alongside the main estimates. Several limitations restrict the conclusions. The primary MR result was borderline significant and had a wide confidence interval. Self-reported insomnia and registry-based AD diagnoses may be misclassified, and the largely European ancestry samples limit generalizability. Reverse

MR results were not presented, leaving a possible genetic effect of AD on insomnia unresolved. The expression datasets differed in platform, tissue region, disease stage, and size; GSE110226 consisted of only 13 choroid plexus samples. ComBat cannot remove all of this biological heterogeneity. Age, sex, brain region, postmortem interval, and other potential confounders were not adequately adjusted for in the combined analysis. The GeneCards set relies on database annotations and may be affected by literature bias. Finally, the selected genes have neither independent external validation nor experimental validation. The results are therefore suited to generating hypotheses and prioritizing candidates, rather than confirming mechanisms.

Future research

The MR analysis needs replication with larger insomnia GWAS using stricter phenotype definitions and with AD datasets covering multiple ancestries. Reverse MR, multivariable MR, and colocalization would help address the direction of effect and the roles of depression, obesity, sleep apnea, and vascular risk factors. Expression analyses should be stratified by brain region. Single-cell or single-nucleus sequencing could resolve cell-specific effects, followed by prespecified testing in independent cohorts. The candidate genes also require quantitative expression and protein-level validation. Cellular and animal experiments could test their effects on inflammation, lipid metabolism, amyloid-β, and tau pathology. Assessing their value as prevention or treatment targets will require agreement between genetic, molecular, and intervention evidence.

Conclusions

Genetic liability to insomnia may be associated with higher late-onset AD risk. The expression analyses suggest neuroinflammation and lipid metabolism as possible molecular links and identify seven genes for further study. Uncertainty in the causal estimate, differences in tissue source, and the lack of independent validation preclude interpreting these genes as established therapeutic targets or clinical diagnostic biomarkers.

List of abbreviations

AD: Alzheimer’s disease; AUC: area under the curve; CI: confidence interval; DEG: differentially expressed gene; FDR: false discovery rate; GEO: Gene Expression Omnibus; GO: Gene Ontology; GWAS: genome-wide association study; ICD-10: International Classification of Diseases, Tenth Revision; IVW: inverse-variance weighted; KEGG: Kyoto Encyclopedia of Genes and Genomes; Lasso: least absolute shrinkage and selection operator; MR: Mendelian randomization; OR: odds ratio; PCA: principal component analysis; PPI: protein–protein interaction; RMA: robust multi-array average; ROC: receiver operating characteristic; SE: standard error; SNP: single-

nucleotide polymorphism; SVM-RFE: support vector machine recursive feature elimination; WGCNA: weighted gene co-expression network analysis

Declarations

Ethics approval and consent to participate

Not applicable. This study was conducted in accordance with the Declaration of Helsinki. It was based exclusively on the secondary analysis of publicly available, de-identified data, namely GWAS summary statistics from the UK Biobank and FinnGen and transcriptomic datasets from the NCBI Gene Expression Omnibus. No human participants were recruited, no biological materials were collected, and no individual-level identifiable information was accessed or analyzed by the authors. Because the study did not involve direct contact with human participants or identifiable personal data, ethical approval and informed consent to participate were not required. Informed consent to participate was obtained from all participants in the original studies by the respective study investigators and data providers, in accordance with the ethical standards of the institutions responsible for those studies.

Consent for publication

Not applicable. This manuscript does not contain data from any individual person.

Availability of data and materials

GWAS summary statistics are available from the UK Biobank (identifier ukb-a-13) and FinnGen (release R13, identifier finn-b-AD_LO). Transcriptomic data are available from the NCBI Gene Expression Omnibus under accession numbers GSE122063 and GSE110226 (https://www.ncbi.nlm.nih.gov/geo/). Insomnia-related genes were retrieved from the GeneCards database (https://www.genecards.org/). Other databases and software versions used in the analyses are listed in the Methods.

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. No funding was obtained for the design of the study, data collection, analysis, interpretation of data, or preparation of the manuscript.

Authors' contributions

YJ conceived and designed the study, performed the Mendelian randomization and bioinformatics analyses, analysed and interpreted the data, and drafted the manuscript. LL acquired and curated the publicly available GWAS summary statistics and GEO datasets, and performed the differential expression and functional enrichment analyses.

XNZ conducted the weighted gene co-expression network analysis and constructed the protein–protein interaction network. XMZ performed the machine learning analyses (Lasso regression and support vector machine recursive feature elimination), evaluated model performance using receiver operating characteristic analysis, and prepared the figures and tables. SQ contributed to the biological interpretation of the results and critically revised the manuscript for important intellectual content. CL supervised the statistical analyses, verified the reproducibility of the analytical code, and revised the manuscript. WW conceived and supervised the project, coordinated the research team, and critically revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the participants and research teams of the UK Biobank, FinnGen, and the contributing GEO studies.

Clinical trial number

Clinical trial number: not applicable.

References

[1] Wang S, Zheng X, Huang J, Liu J, Li C, Shang H. Sleep characteristics and risk of Alzheimer's disease: a systematic review and meta-analysis of longitudinal studies. J Neurol. 2024;271:3782–3793. doi:10.1007/s00415-024-12380-7.

[2] Meng M, Zhang Y, Li J, et al. Insomnia and risk of all-cause dementia: a systematic review and meta-analysis. PLoS One. 2025;20(4):e0318814. doi:10.1371/journal.pone.0318814.

[3] Lacerda RAV, Desio JAF, Kammers CM, et al. Sleep disorders and risk of Alzheimer's disease: a two-way road. Ageing Res Rev. 2024;101:102514. doi:10.1016/j.arr.2024.102514.

[4] Chong PLH, Garic D, Shen MD, Lundgaard I, Schwichtenberg AJ. Sleep, cerebrospinal fluid, and the glymphatic system: a systematic review. Sleep Med Rev. 2022;61:101572. doi:10.1016/j.smrv.2021.101572.

[5] Zhang Y, Ren R, Yang L, et al. Sleep in Alzheimer's disease: a systematic review and meta-analysis of polysomnographic findings. Transl Psychiatry. 2022;12:136. doi:10.1038/s41398-022-01897-y.

[6] Anderson EL, Richmond RC, Jones SE, et al. Is disrupted sleep a risk factor for Alzheimer's disease? Evidence from a two-sample Mendelian randomization analysis. Int J Epidemiol. 2021;50(3):817–828. doi:10.1093/ije/dyaa183.

[7] Skrivankova VW, Richmond RC, Woolf BAR, et al. Strengthening the reporting of observational studies in epidemiology using Mendelian randomization: the STROBE-MR statement. JAMA. 2021;326(16):1614–1621. doi:10.1001/jama.2021.18236.

[8] Lane JM, Jones SE, Dashti HS, et al. Biological and clinical insights from genetics of insomnia symptoms. Nat Genet. 2019;51(3):387–393. doi:10.1038/s41588-019-0361-7.

[9] Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015;44(2):512–525. doi:10.1093/ije/dyv080.

[10] Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epidemiol. 2016;40(4):304–314. doi:10.1002/gepi.21965.

[11] Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. Int J Epidemiol. 2017;46(6):1985–1998. doi:10.1093/ije/dyx102.

[12] Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. eLife. 2018;7:e34408. doi:10.7554/eLife.34408.

[13] Stopa EG, Tanis KQ, Miller MC, et al. Comparative transcriptomics of choroid plexus in Alzheimer's disease, frontotemporal dementia and Huntington's disease: implications for CSF homeostasis. Fluids Barriers CNS. 2018;15:18. doi:10.1186/s12987-018-0102-9.

[14] Clough E, Barrett T. The Gene Expression Omnibus database. Methods Mol Biol. 2016;1418:93–110. doi:10.1007/978-1-4939-3578-9_5.

[15] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–127. doi:10.1093/biostatistics/kxj037.

[16] Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. doi:10.1093/nar/gkv007.

[17] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

[18] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. doi:10.1186/1471-2105-9-559.

[19] Szklarczyk D, Kirsch R, Koutrouli M, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced

genome of interest. Nucleic Acids Res. 2023;51(D1):D638–D646. doi:10.1093/nar/gkac1000.

[20] Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–2504. doi:10.1101/gr.1239303.

[21] Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2(3):100141. doi:10.1016/j.xinn.2021.100141.

[22] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22. doi:10.18637/jss.v033.i01.

[23] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389–422. doi:10.1023/A:1012487302797.

[24] Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77. doi:10.1186/1471-2105-12-77.

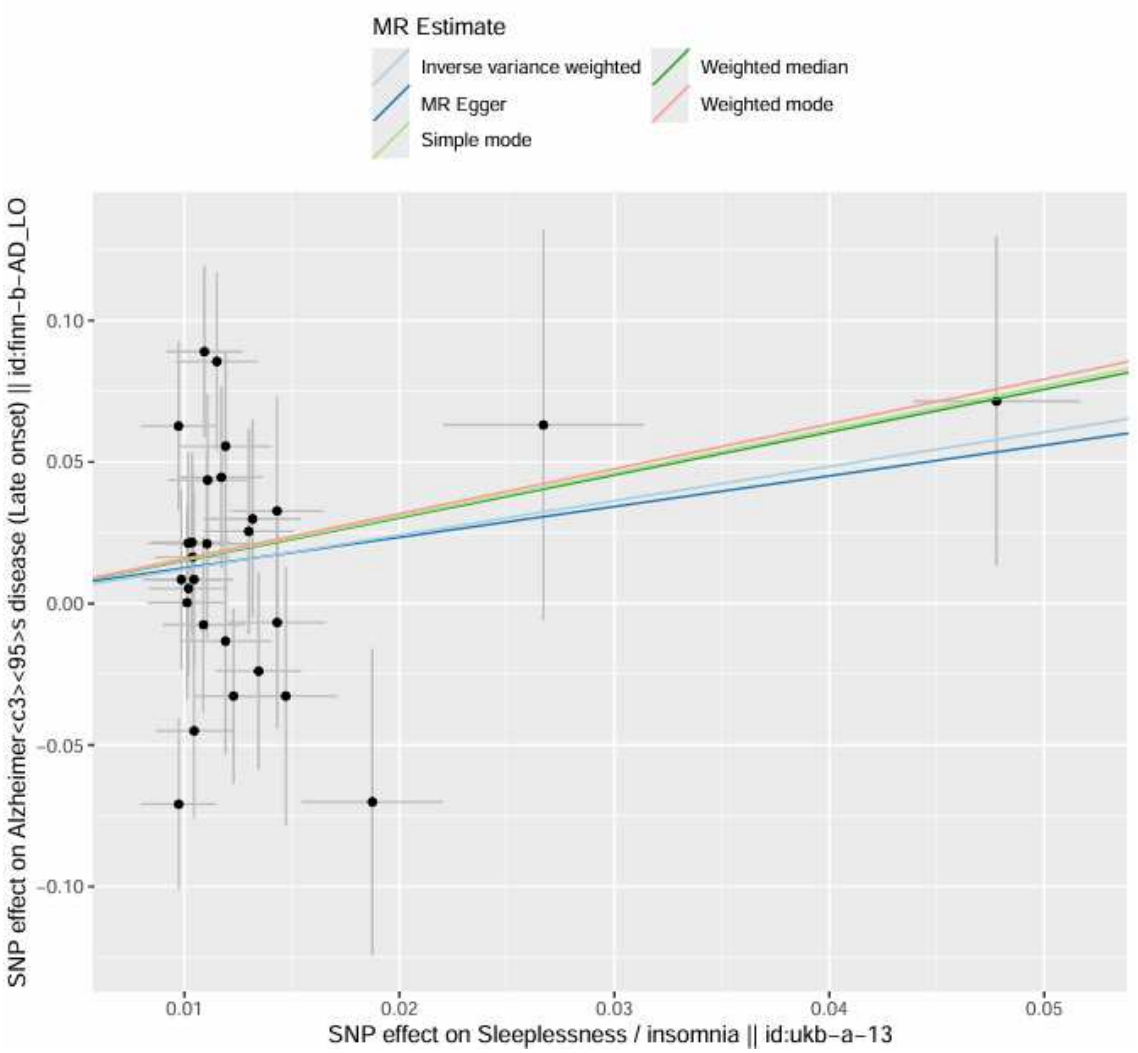
[25] Leng F, Edison P. Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? Nat Rev Neurol. 2021;17(3):157–172. doi:10.1038/s41582-020-00435-y.

[26] Yin F. Lipid metabolism and Alzheimer's disease: clinical evidence, mechanistic link and therapeutic promise. FEBS J. 2023;290(6):1420–1453. doi:10.1111/febs.16344.

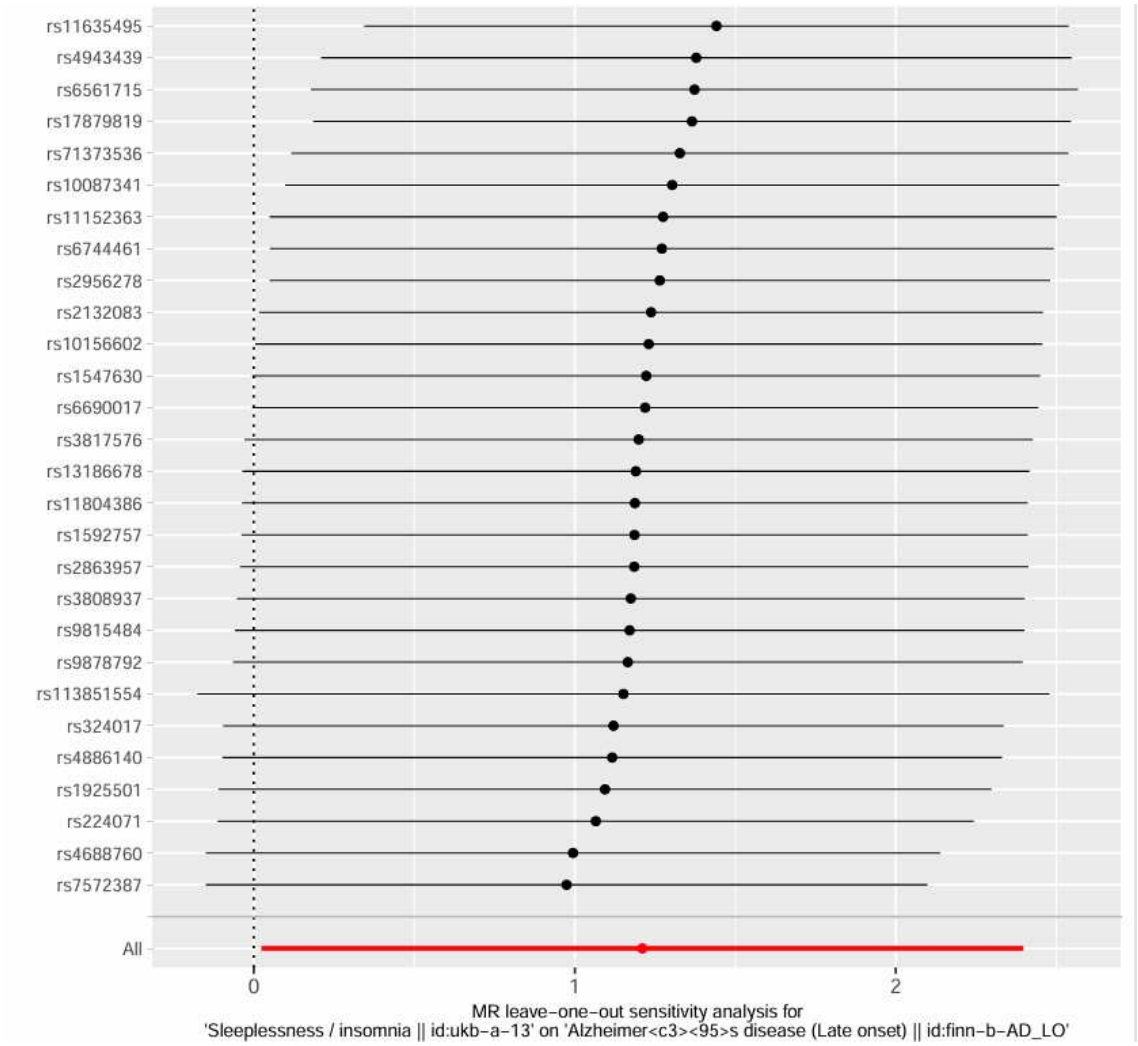
Figure legends

Figure 1. Mendelian randomization analysis of the causal effect of insomnia on late-onset Alzheimer's disease. (A) Scatter plot of SNP effects on insomnia (x-axis) and on Alzheimer's disease (y-axis); the five regression lines represent the estimates obtained with inverse-variance weighting, MR-Egger, simple mode, weighted median, and weighted mode methods, all indicating a positive causal association. (B) Leave-one-out analysis showing that the overall estimate is not driven by any single SNP. (C) Funnel plot showing the approximately symmetrical distribution of individual SNP estimates around the overall effect, suggesting no significant pleiotropy. (D) Forest plot of the MR-Egger and inverse-variance weighted estimates with 95% confidence intervals.

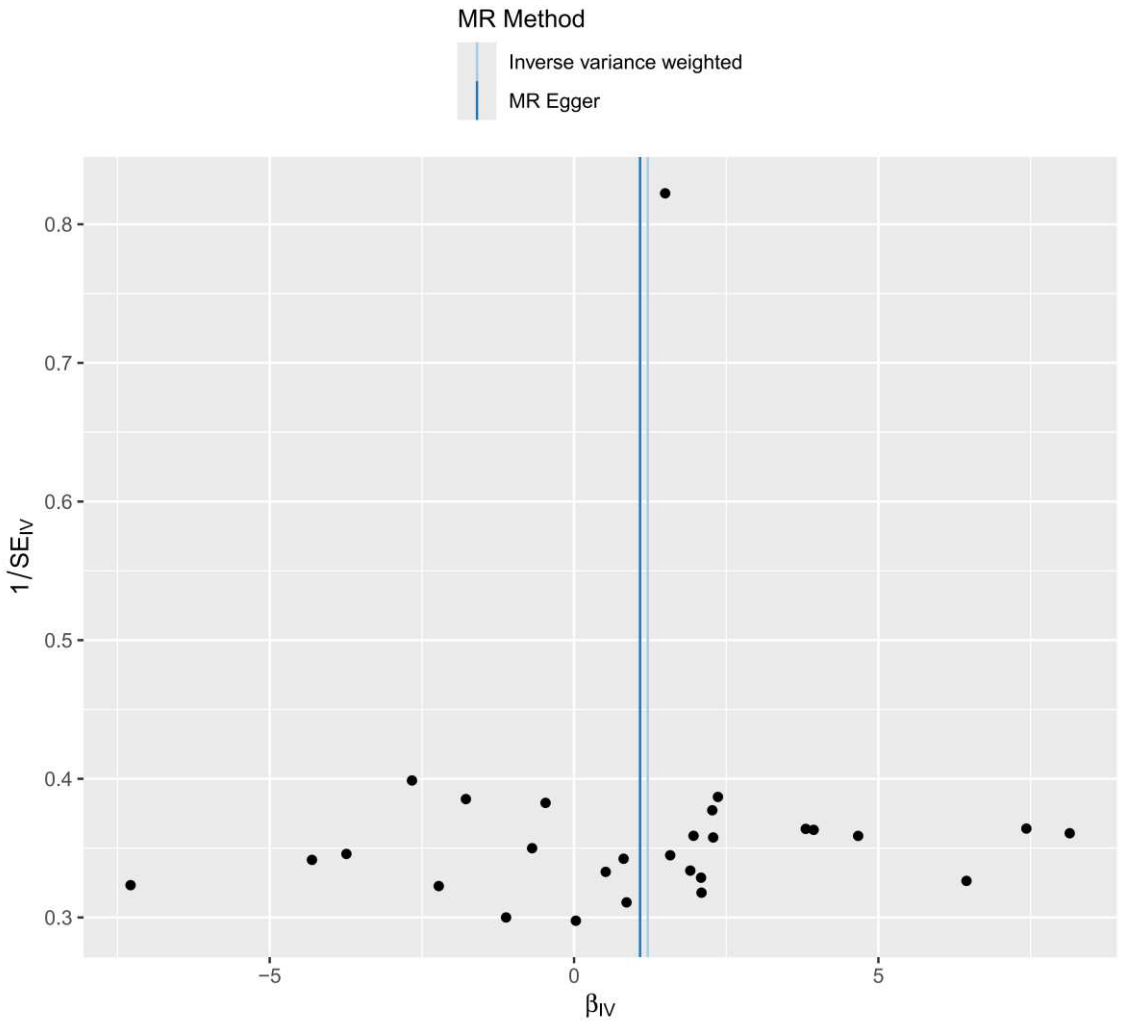
(A)


MR Estimate
Inverse variance weighted
MR Egger
Simple mode
Weighted median
Weighted mode
SNP effect on Alzheimer<c3><95>s disease (Late onset) || id:finn-b-AD_LO
0.10
0.05
0.00
-0.05
-0.10
0.01
0.02
0.03
0.04
0.05
SNP effect on Sleeplessness / insomnia || id:ukb-a-13

(B)


rs11635495
rs4943439
rs6561715
rs17879819
rs71373536
rs10087341
rs11152363
rs6744461
rs2956278
rs2132083
rs10156602
rs1547630
rs6690017
rs3817576
rs13186678
rs11804386
rs1592757
rs2863957
rs3808937
rs9815484
rs9878792
rs113851554
rs324017
rs4886140
rs1925501
rs224071
rs4688760
rs7572387
All
0
1
2
MR leave-one-out sensitivity analysis for
'Sleeplessness / insomnia || id:ukb-a-13' on 'Alzheimer<c3><95>s disease (Late onset) || id:finn-b-AD_LO'

(C)


MR Method
Inverse variance weighted
MR Egger
0.8
0.7
0.6
0.5
0.4
0.3
1/SE_IV
−5
0
5
β_IV

(D)

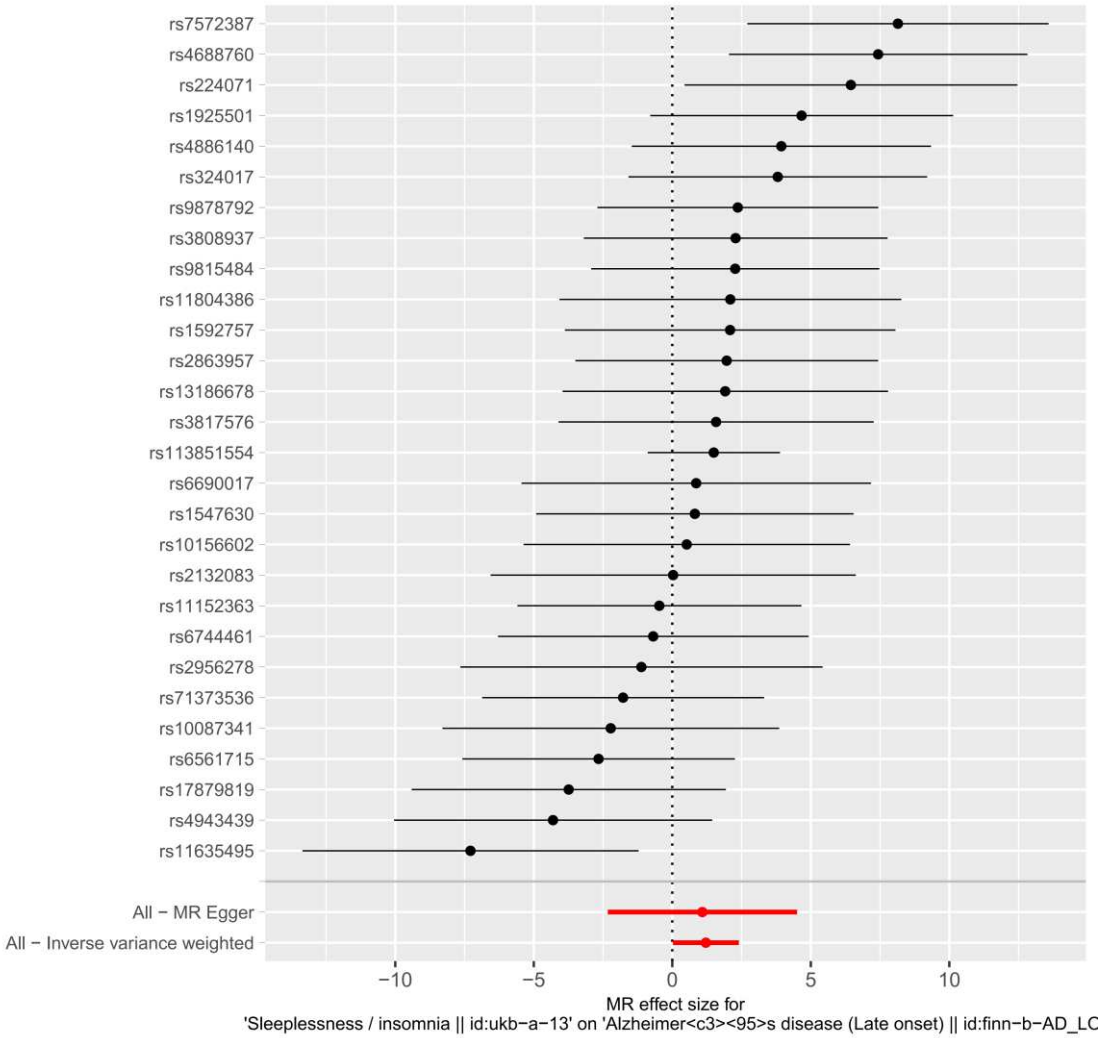
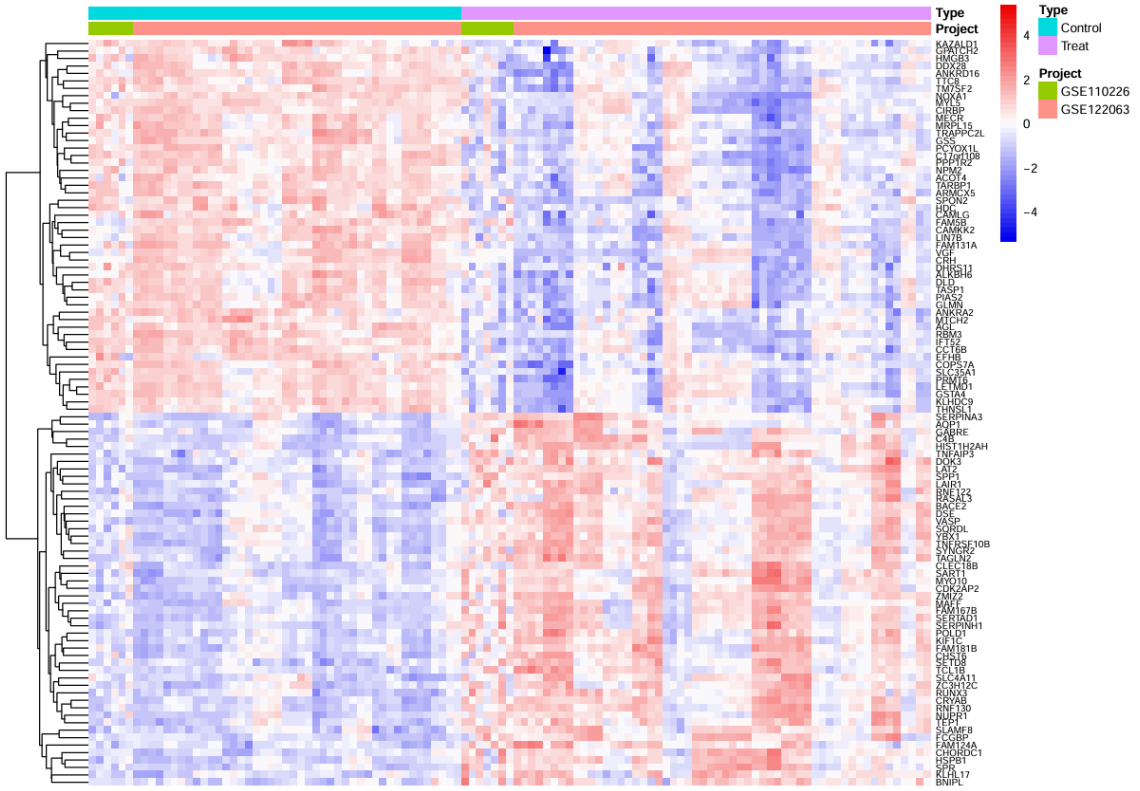


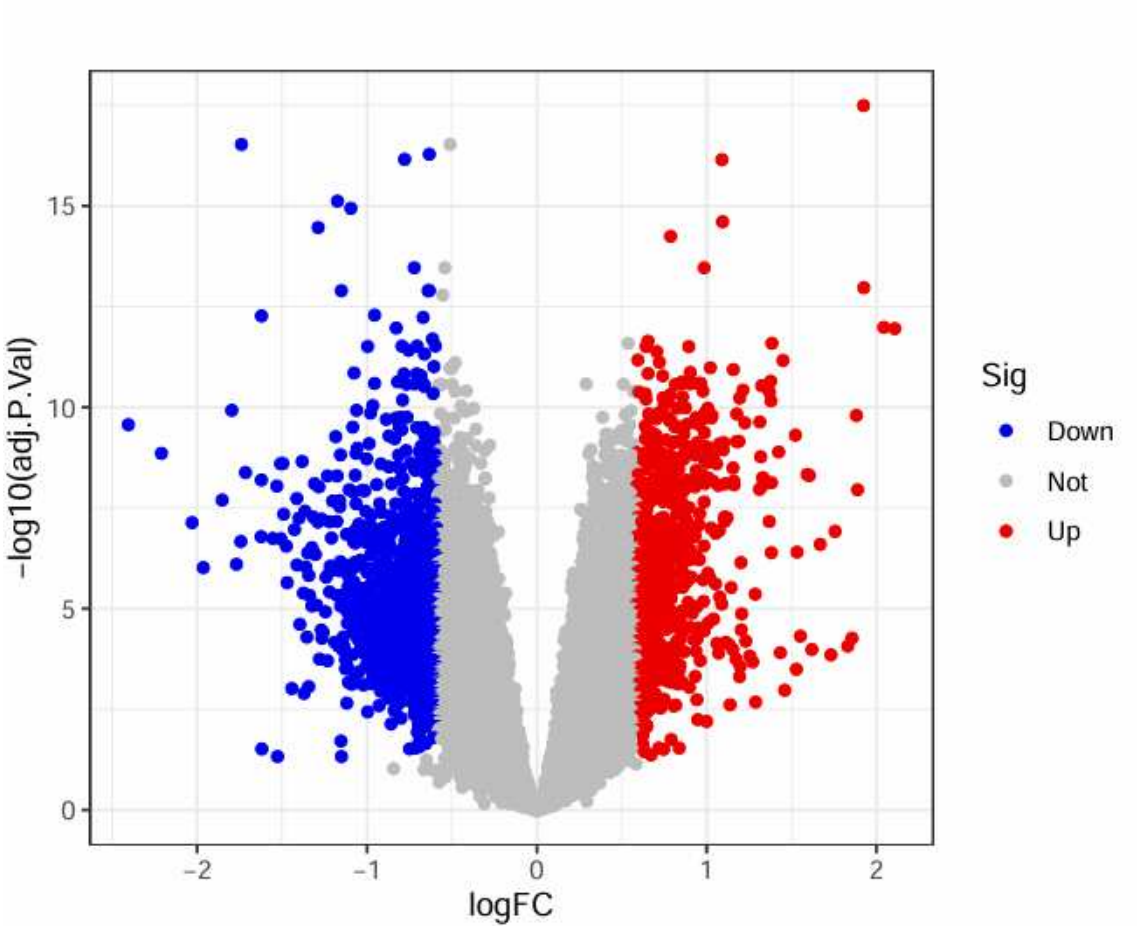
Figure 2. Transcriptomic profiling and bioinformatic analyses of the molecular mechanisms linking insomnia and Alzheimer's disease. (A) Heatmap of the 50 most significantly differentially expressed genes between AD and control samples in the combined GSE122063 and GSE110226 dataset after ComBat batch correction. (B) Volcano plot of differentially expressed genes ($|\log_2 FC| > 0.58$, FDR < 0.05). (C) Sample dendrogram with the clinical trait heatmap of the 113 combined samples, showing no outliers requiring exclusion. (D) Gene dendrogram obtained by hierarchical clustering; the color band below indicates the eight co-expression modules identified by WGCNA. (E) Module–trait relationship heatmap showing the Pearson correlations between module eigengenes and AD status; correlation coefficients and P values are given in each cell. (F) Venn diagram of insomnia-related genes (GeneCards), differentially expressed genes (GEO), and genes in the AD-associated WGCNA modules, yielding 35 candidate genes. (G) Protein–protein interaction network of the candidate genes constructed with STRING (confidence score > 0.7) and visualized in Cytoscape. (H) Circular plot of enriched Gene Ontology terms across biological process, cellular component, and molecular function

categories. (I) Bubble plot of the top enriched Gene Ontology terms in the biological process (BP), cellular component (CC), and molecular function (MF) categories. (J) Bubble plot of the enriched KEGG pathways, including the Influenza A and Lipid and atherosclerosis pathways.

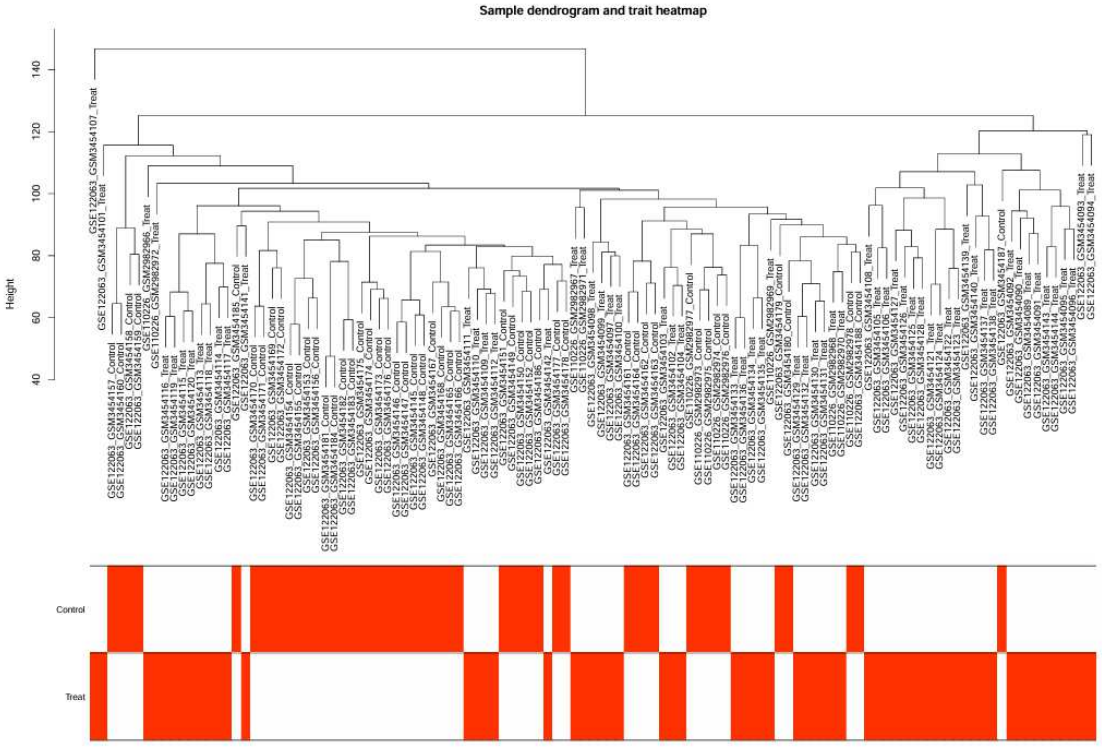
(A)



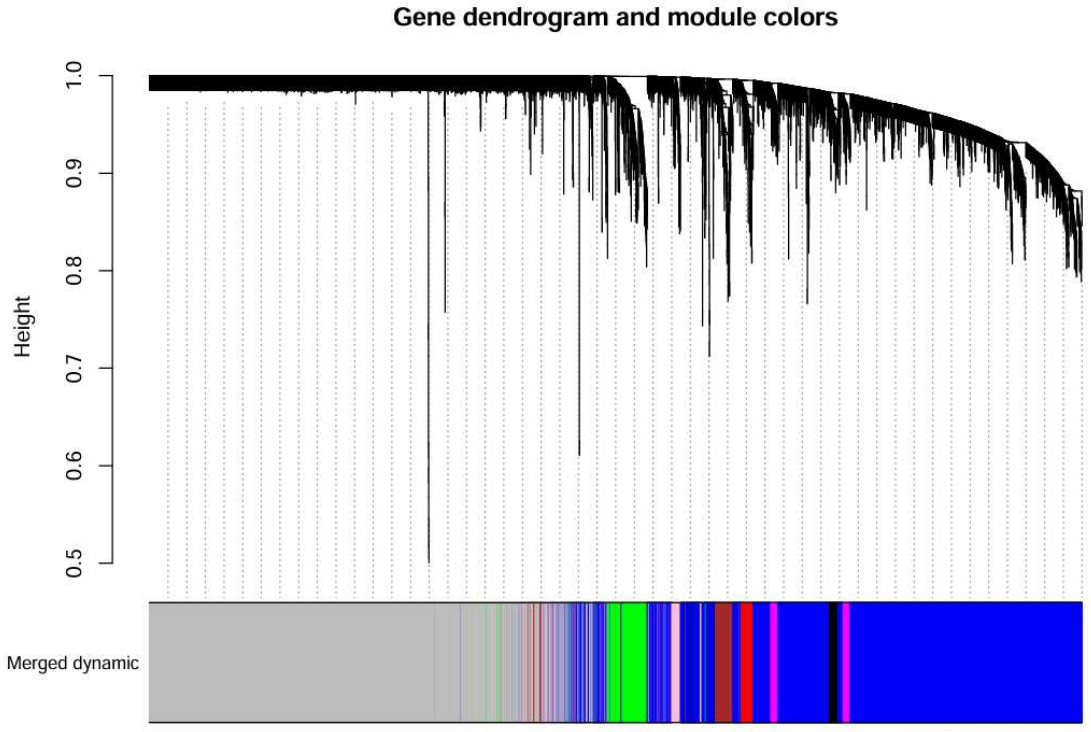
(B)



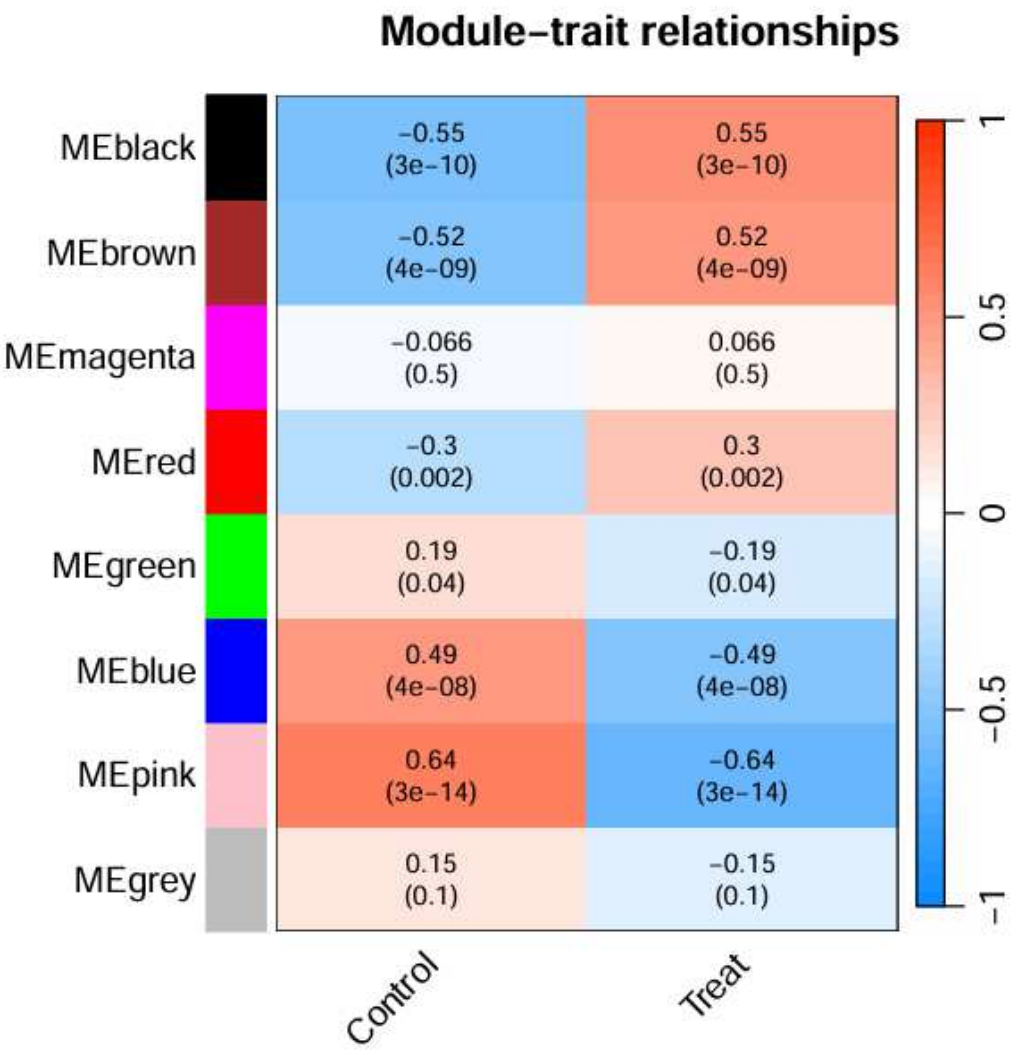
(C)


Sample dendrogram and trait heatmap
Height
Control
Treat

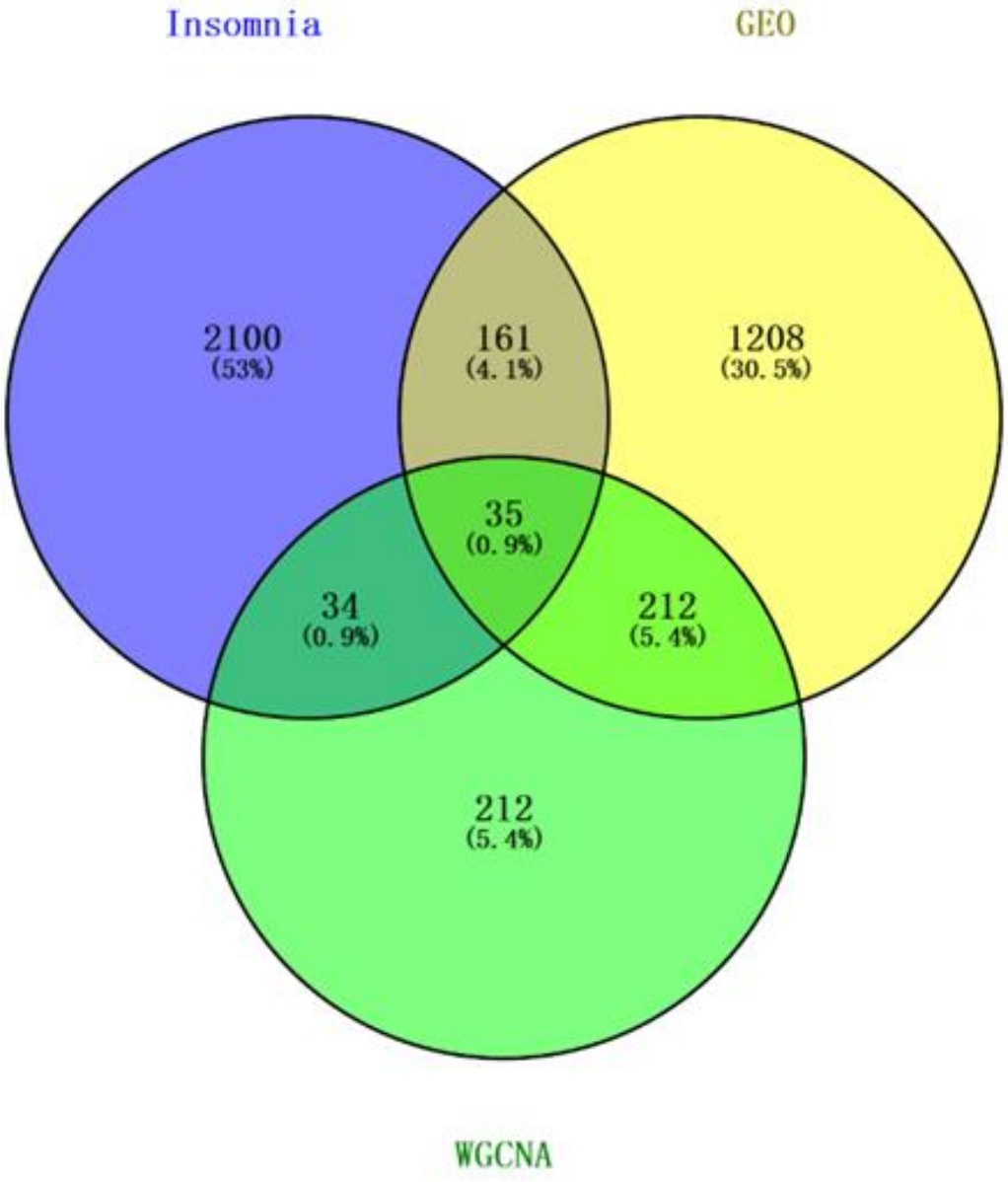
(D)


Gene dendrogram and module colors
Height
Merged dynamic

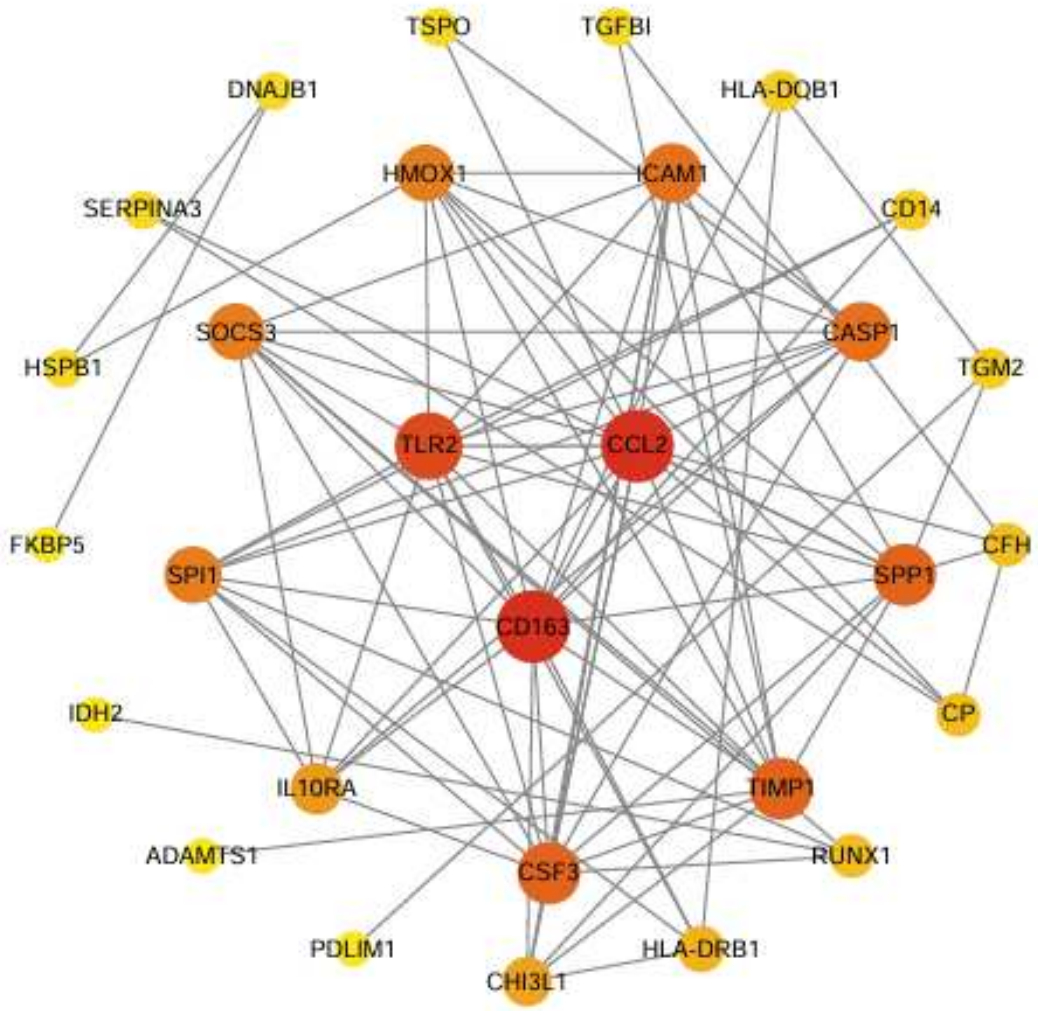
(E)


Module−trait relationships
MEblack
MEbrown
MEmagenta
MEred
MEgreen
MEblue
MEpink
MEgrey
−0.55
(3e−10)
0.55
(3e−10)
−0.52
(4e−09)
0.52
(4e−09)
−0.066
(0.5)
0.066
(0.5)
−0.3
(0.002)
0.3
(0.002)
0.19
(0.04)
−0.19
(0.04)
0.49
(4e−08)
−0.49
(4e−08)
0.64
(3e−14)
−0.64
(3e−14)
0.15
(0.1)
−0.15
(0.1)
Control
Treat
1
0.5
0
−0.5
−1

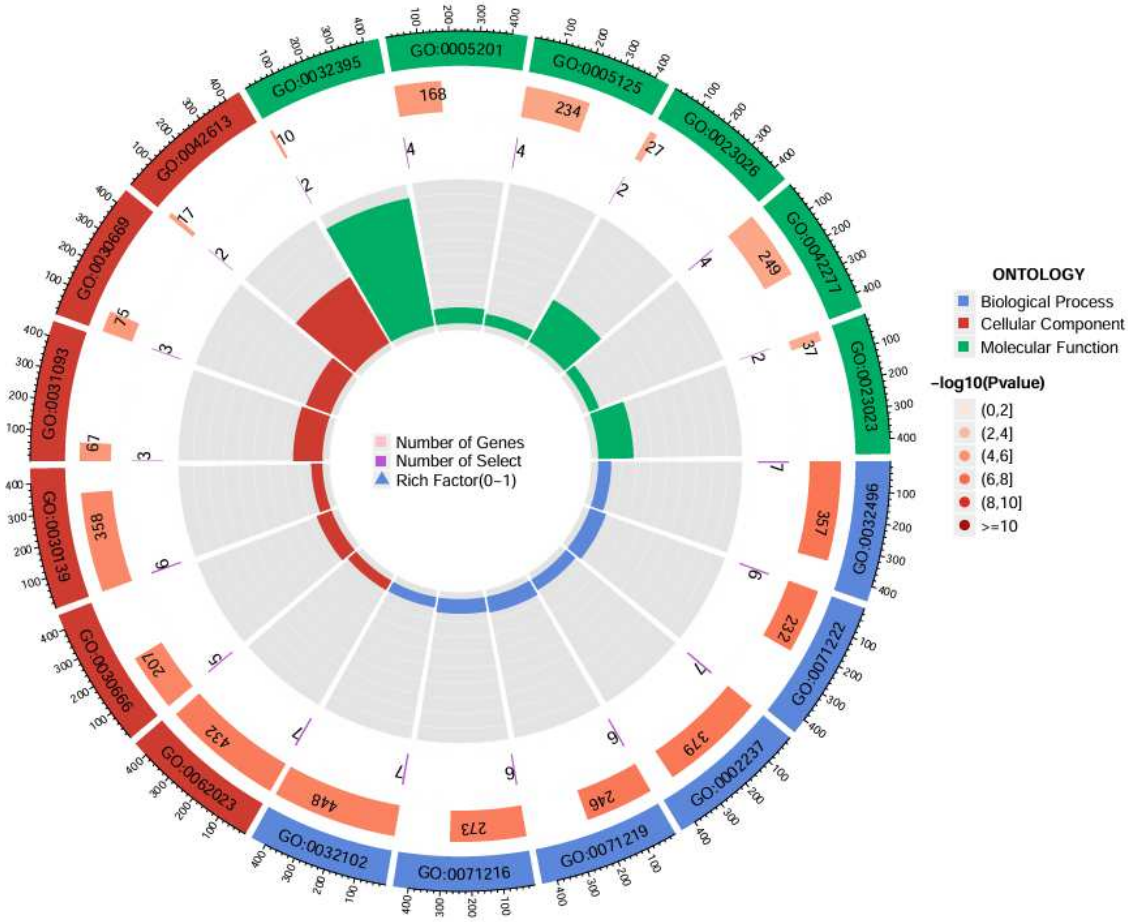
(F)


Insomnia
GEO
2100
(53%)
161
(4.1%)
1208
(30.5%)
35
(0.9%)
34
(0.9%)
212
(5.4%)
212
(5.4%)
WGCNA

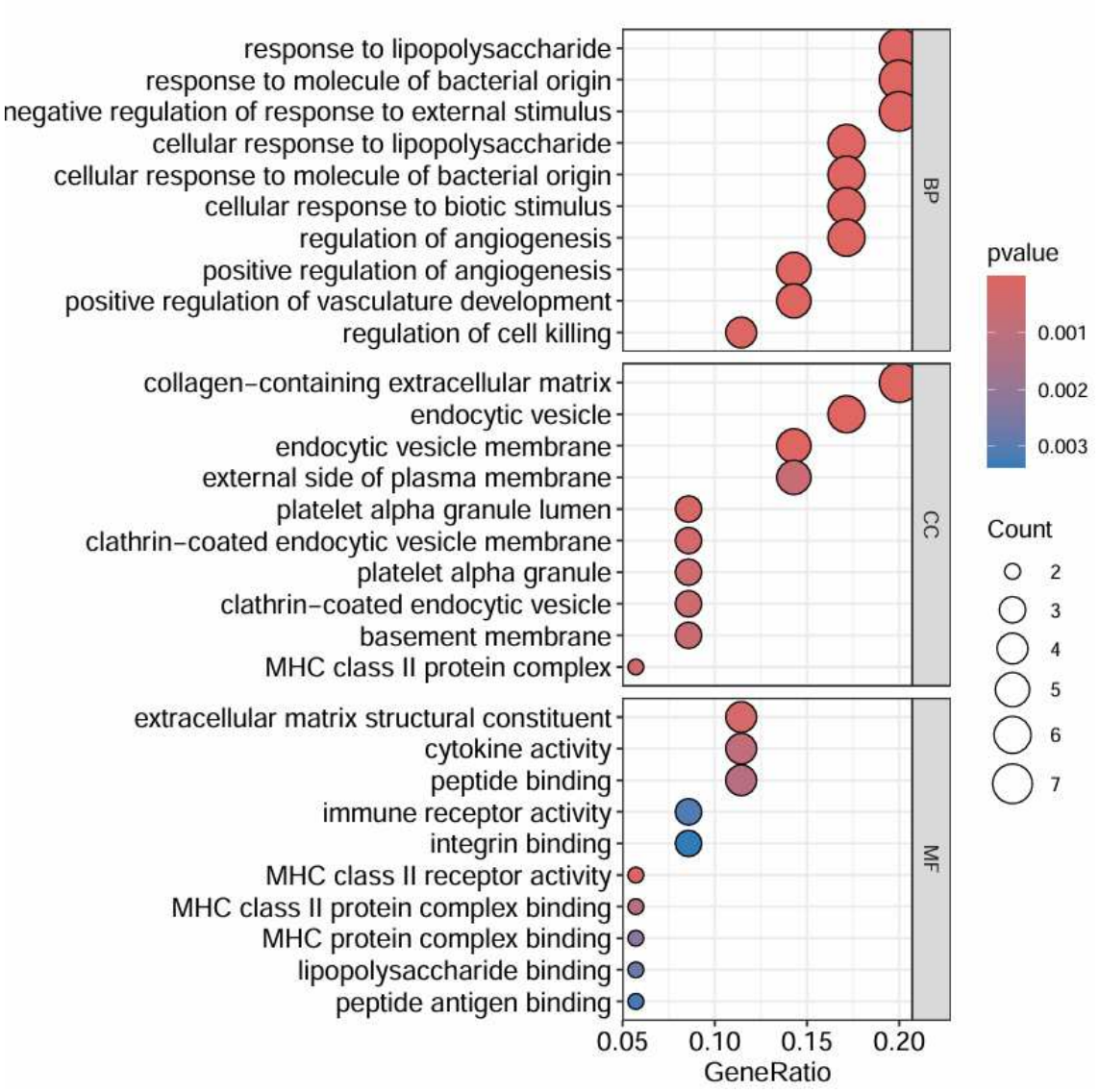
(G)


TSPO
TGFBI
DNAJB1
HLA-DQB1
HMOX1
ICAM1
SERPINA3
CD14
SOCS3
CASP1
HSPB1
TGM2
TLR2
CCL2
FKBP5
CFH
SPI1
SPP1
CD163
IDH2
CP
IL10RA
TIMP1
ADAMTS1
CSF3
RUNX1
PDLIM1
HLA-DRB1
CHI3L1

(H)


GO:0005201
GO:0005125
GO:0023026
GO:0042277
GO:0023023
GO:0032496
GO:0071222
GO:0002237
GO:0071219
GO:0071216
GO:0032102
GO:0062023
GO:0030666
GO:0030139
GO:0031093
GO:0030669
GO:0042613
GO:0032395
168
234
27
249
37
357
232
379
246
273
448
432
207
358
67
75
17
10
Number of Genes
Number of Select
Rich Factor(0-1)
ONTOLOGY
Biological Process
Cellular Component
Molecular Function
-log10(Pvalue)
(0,2]
(2,4]
(4,6]
(6,8]
(8,10]
>=10

(I)


response to lipopolysaccharide
response to molecule of bacterial origin
negative regulation of response to external stimulus
cellular response to lipopolysaccharide
cellular response to molecule of bacterial origin
cellular response to biotic stimulus
regulation of angiogenesis
positive regulation of angiogenesis
positive regulation of vasculature development
regulation of cell killing
BP
collagen−containing extracellular matrix
endocytic vesicle
endocytic vesicle membrane
external side of plasma membrane
platelet alpha granule lumen
clathrin−coated endocytic vesicle membrane
platelet alpha granule
clathrin−coated endocytic vesicle
basement membrane
MHC class II protein complex
CC
extracellular matrix structural constituent
cytokine activity
peptide binding
immune receptor activity
integrin binding
MHC class II receptor activity
MHC class II protein complex binding
MHC protein complex binding
lipopolysaccharide binding
peptide antigen binding
MF
0.05
0.10
0.15
0.20
GeneRatio
pvalue
0.001
0.002
0.003
Count
2
3
4
5
6
7

(J)

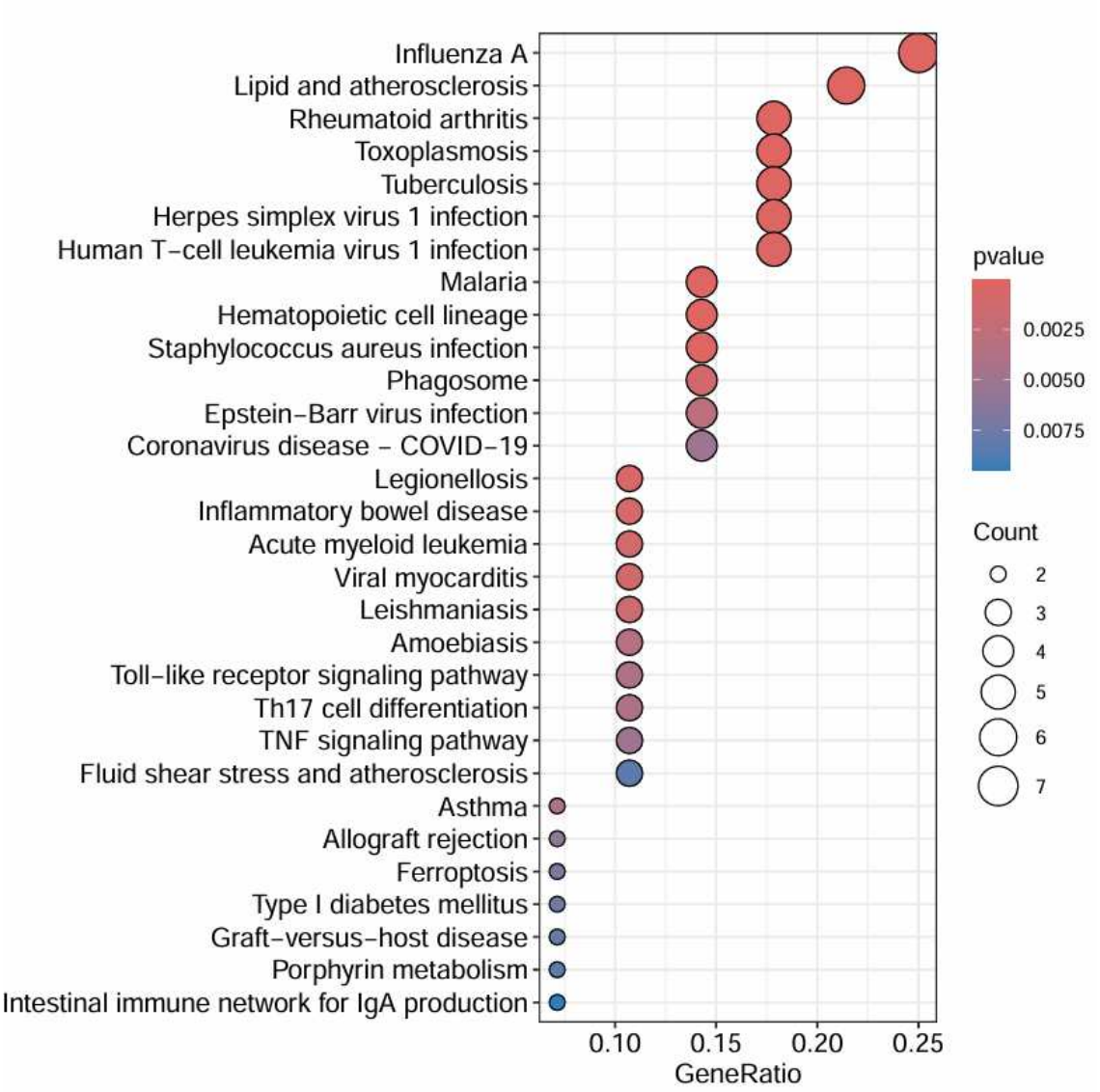
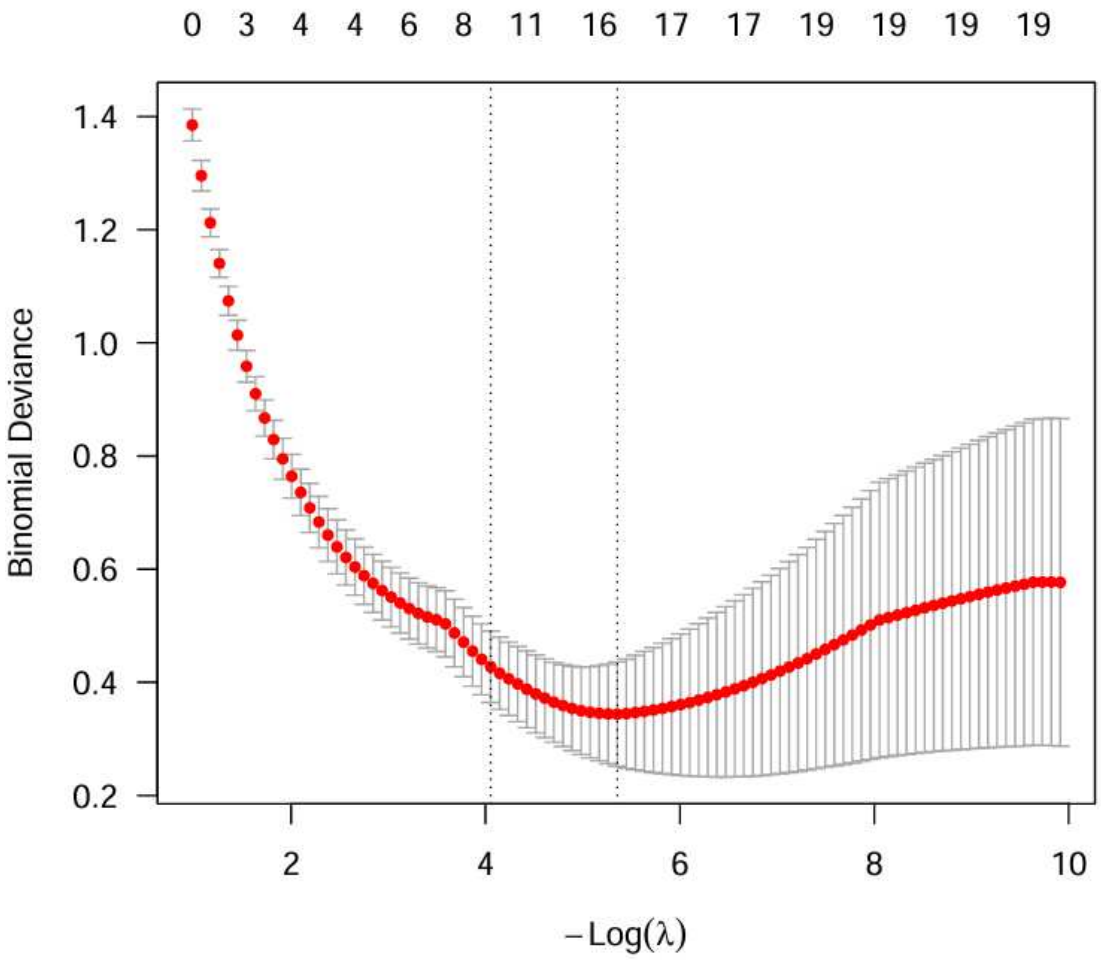
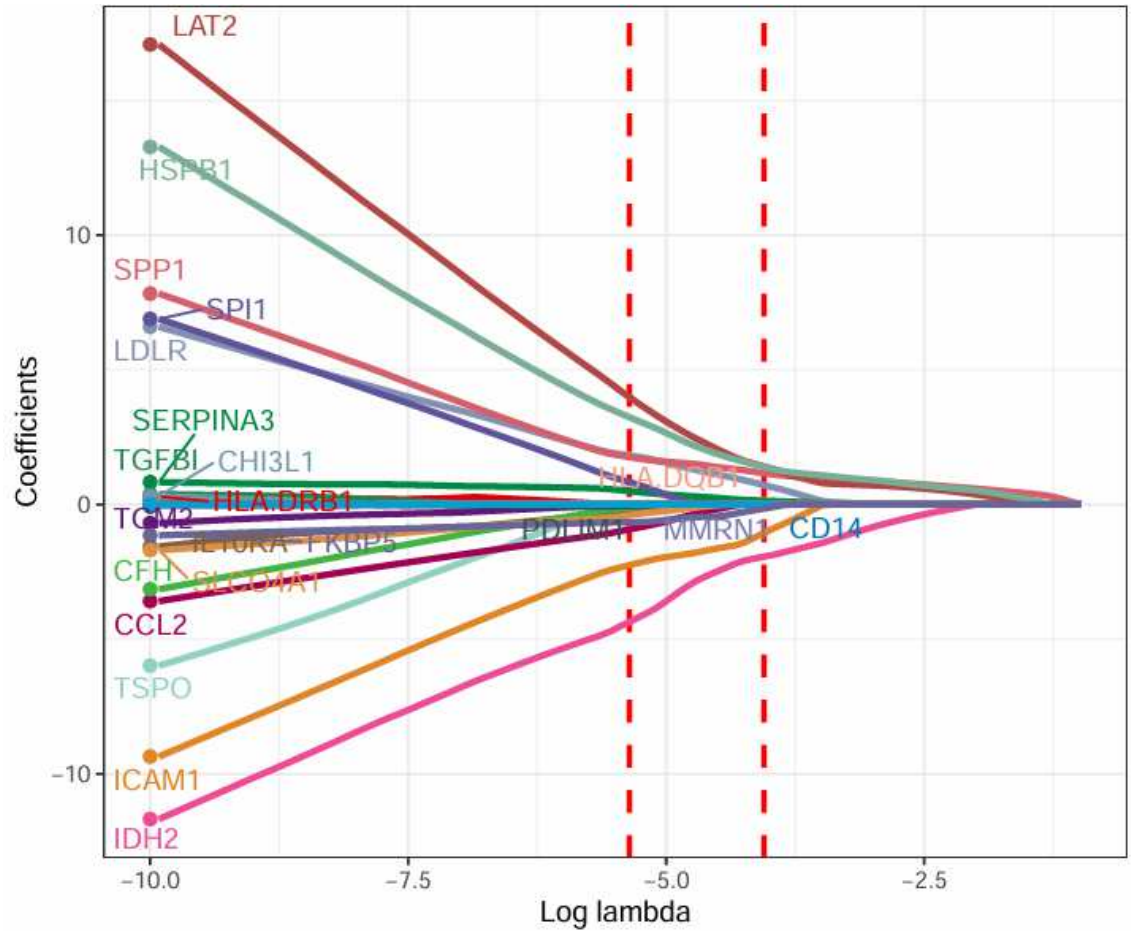


Figure 3. Machine learning selection of key target genes and assessment of diagnostic performance. (A) Cross-validation curve of Lasso regression showing binomial deviance across λ values; dashed lines indicate the optimal λ. (B) Lasso coefficient profile plot showing the shrinkage of gene coefficients as λ changes; the red dashed lines mark the optimal λ. (C) Support vector machine recursive feature elimination (SVM-RFE) curve relating the number of features to the cross-validated accuracy; the circle marks the optimal number of features. (D) Venn diagram of the genes selected by SVM-RFE and Lasso, identifying seven key genes. (E) Box plots showing the expression of the seven key genes in AD samples (red) and controls (blue) in the combined GEO dataset (***P < 0.001). (F) Receiver operating characteristic curves for the seven key genes with the corresponding AUC values.

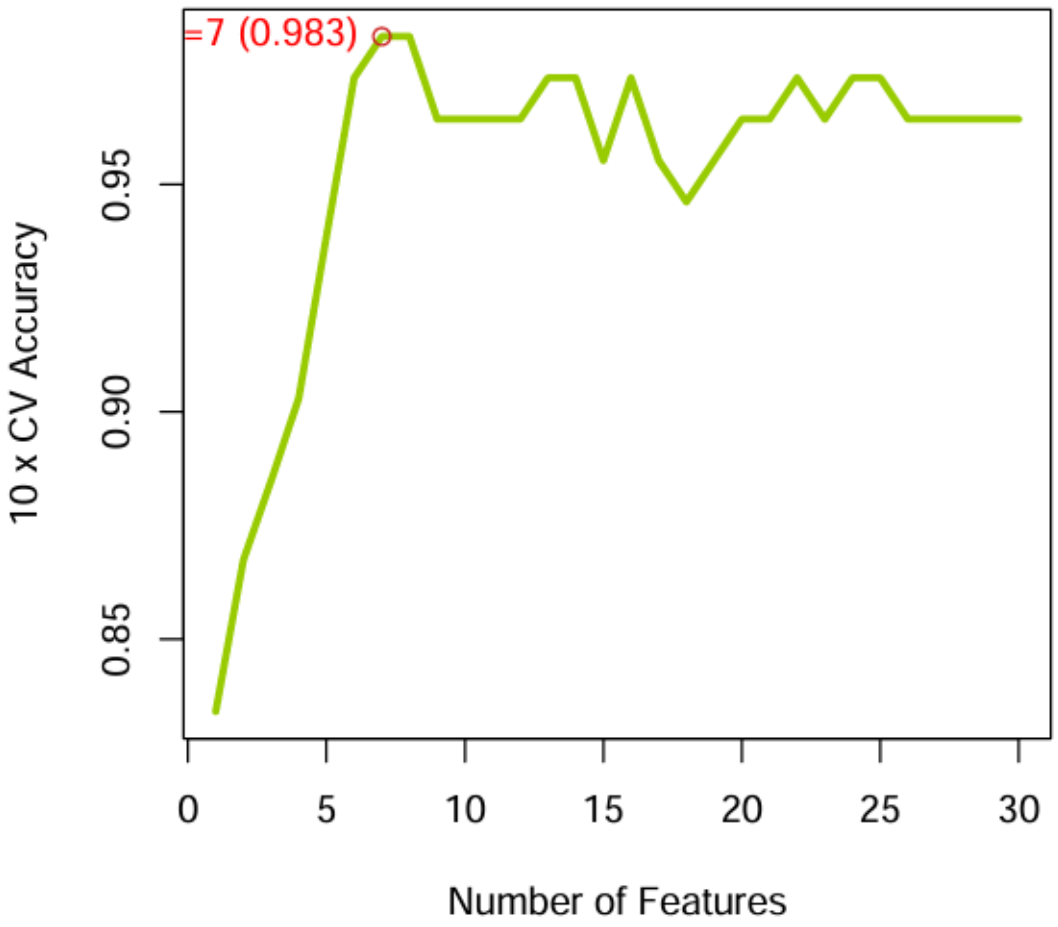
(A)


0 3 4 4 6 8 11 16 17 17 19 19 19 19
Binomial Deviance
1.4
1.2
1.0
0.8
0.6
0.4
0.2
2
4
6
8
10
−Log(λ)

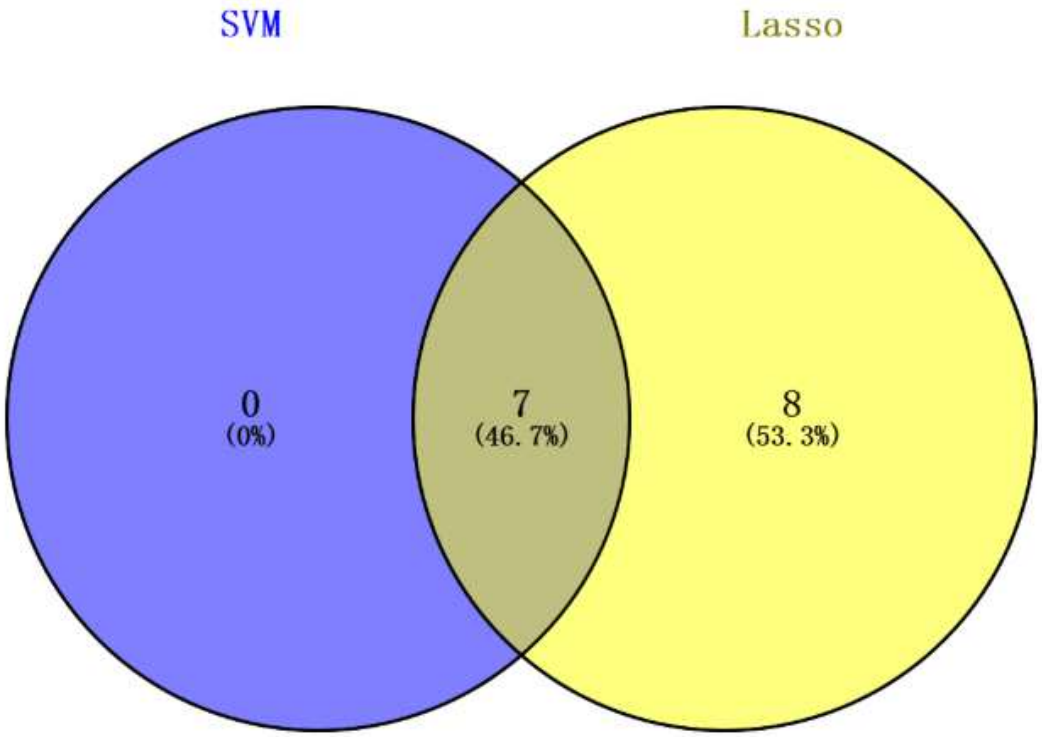
(B)


LAT2
HSPB1
SPP1
SPI1
LDLR
SERPINA3
TGFBI
CHI3L1
HLA.DRB1
TGM2
CFH
CCL2
TSPO
ICAM1
IDH2
HLA.DQB1
PDLIM1
MMRN1
CD14
10
0
−10
Coefficients
−10.0
−7.5
−5.0
−2.5
Log lambda

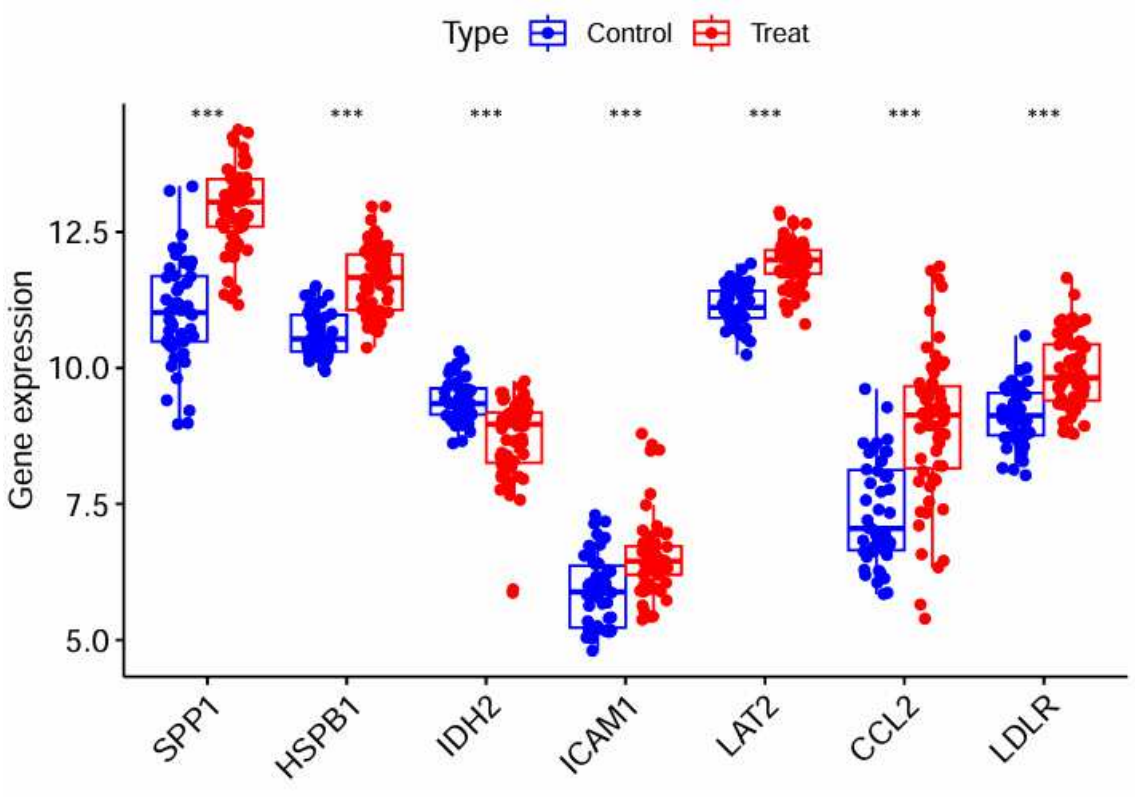
(C)


=7 (0.983)
10 x CV Accuracy
0.85
0.90
0.95
0
5
10
15
20
25
30
Number of Features

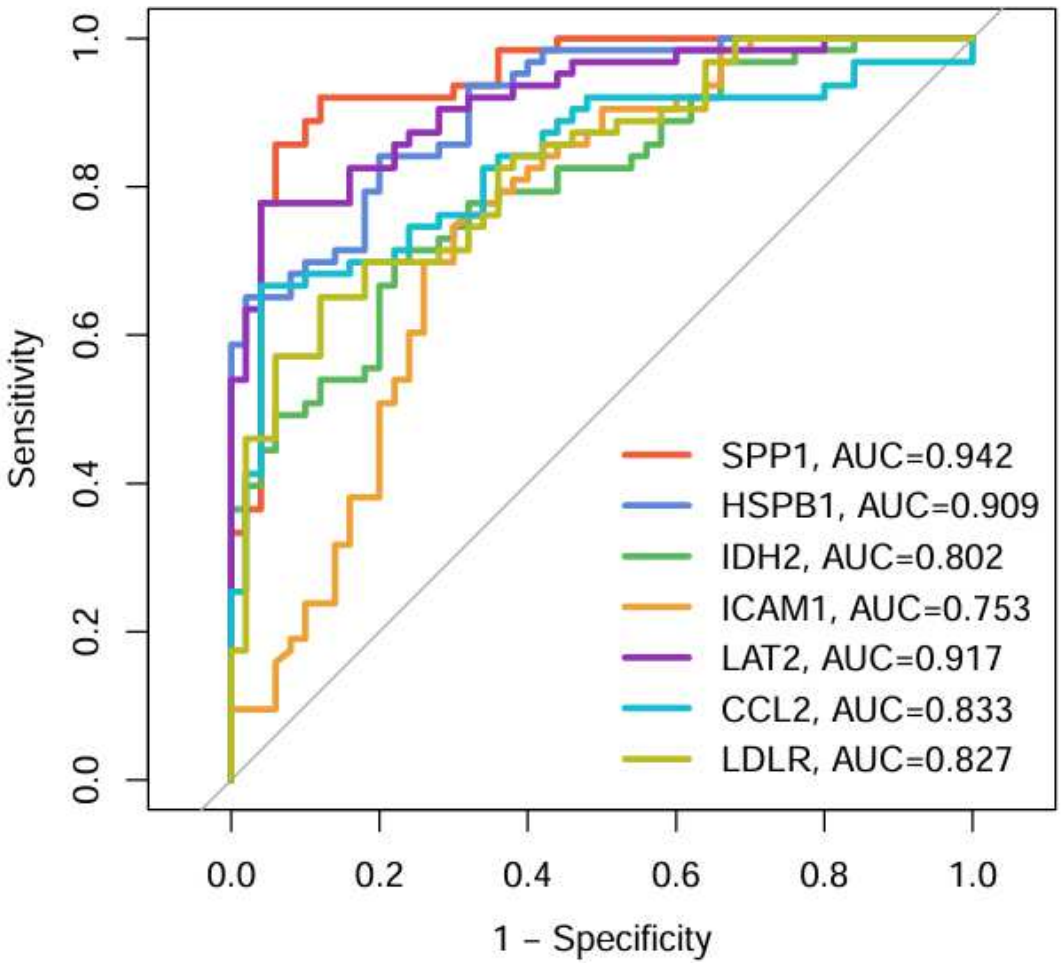
(D)


SVM
Lasso
0
(0%)
7
(46.7%)
8
(53.3%)

(E)


Type
Control
Treat
Gene expression
12.5
10.0
7.5
5.0
***
SPP1
HSPB1
IDH2
ICAM1
LAT2
CCL2
LDLR

(F)



Tables

Table 1. Mendelian randomization estimates for the causal effect of insomnia on late-onset Alzheimer's disease.

Method	**SNPs**	**OR**	**95% CI**	**P**
MR Egger	28	2.96	0.10–89.77	0.54
Weighted median	28	4.54	0.95–21.70	0.06
IVW	28	3.36	1.02–10.99	0.045
Simple mode	28	4.66	0.29–74.66	0.29
Weighted mode	28	4.88	0.58–40.81	0.15