nature communications

Article

https://doi.org/10.1038/s41467-025-59949-y

Whole-genome sequencing analyses suggest novel genetic factors associated with Alzheimer's disease and a cumulative effects model for risk liability

Received: 26 May 2024

Accepted: 8 May 2025

Published online: 26 May 2025

Check for updates

Jun Pyo Kim[1,2,3,21], Minyoung Cho[4,21], Chanhee Kim[5,6,21], Hyunwoo Lee[7], Beomjin Jang[4,8,9,10], Sang-Hyuk Jung[11], Yujin Kim[5,6], In Gyeong Koh[5,6], Seoyeon Kim[5,6], Daeun Shin[1,2,3], Eun Hye Lee[2,12,13], Jong-Young Lee[14], YoungChan Park[15], Hyemin Jang[1,16], Bo-Hyun Kim[1], Hongki Ham[2], Beomsu Kim[4], Yujin Kim[4], A-Hyun Cho[4], Towfique Raj[8,9,10,17,18], Hee Jin Kim[1,2,3], Duk L. Na[1,2,3], Sang Won Seo[1,2,3,7,22] ✉, Joon-Yong An[5,6,19,22] ✉ & Hong-Hee Won[4,20,22] ✉

Genome-wide association studies (GWAS) on Alzheimer's disease (AD) have predominantly focused on identifying common variants in Europeans. Here, we performed whole-genome sequencing (WGS) of 1,559 individuals from a Korean AD cohort to identify various genetic variants and biomarkers associated with AD. Our GWAS analysis identified a previously unreported locus for common variants (*APCDD1*) associated with AD. Our WGS analysis was extended to explore the less-characterized genetic factors contributing to AD risk. We identified rare noncoding variants located in cis-regulatory elements specific to excitatory neurons associated with cognitive impairment. Moreover, structural variation analysis showed that short tandem repeat expansion was associated with an increased risk of AD, and copy number variant at the *HPSE2* locus showed borderline statistical significance. *APOE* ε4 carriers with high polygenic burden or structural variants exhibited severe cognitive impairment and increased amyloid beta levels, suggesting a cumulative effects model of AD risk.

Alzheimer's disease (AD) is the leading cause of dementia worldwide, affecting over 57 million individuals[1]. AD is a complex neurodegenerative disorder influenced by genetic and environmental factors, with approximately 60–80% heritability[2]. Recently, large-scale genome-wide association studies (GWAS) have discovered 75 AD-associated genetic loci[3,4]; however, these only account for approximately 15% of the phenotypic variance[5], indicating that a substantial portion of the genetic factors involved in AD remain to be discovered. Despite these findings, most previous GWAS have been conducted primarily in European populations, highlighting the need for research in diverse populations. Indeed, with only a small fraction of genetic variants shared across all ancestries[6], GWAS of diverse populations holds the potential to identify novel genetic factors[7] and population-specific rare variants[8–11].

Whole-genome sequencing (WGS) analysis can identify the full spectrum of genetic factors and facilitate genetic association testing for common and rare variants, thereby elucidating the genetic architecture of AD. WGS evaluates the contribution of rare noncoding variations, which are largely unexplored but constitute a significant proportion of an individual's genetic variation. The noncoding genome

A full list of affiliations appears at the end of the paper. ✉e-mail: sangwonseo@empas.com; joonan30@korea.ac.kr; wonhh@skku.edu

harbors regulatory elements such as enhancers and promoters, which are essential for cell-type-specific gene expression and cellular function. Rare noncoding variants within these elements can disrupt gene regulatory networks and potentially affect AD development and progression[12]. In addition, WGS analysis can identify structural variants (SVs) such as copy number variations (CNVs) and short tandem repeats (STRs), which previous studies have rarely investigated simultaneously. A large-scale WGS study reported a rare AD-associated STR in the downstream noncoding region of *APOE*[13].

Although previous large-scale GWAS have used clinical diagnosis as an outcome, core AD biomarkers are critical for genetic studies because patients clinically diagnosed with AD may not exhibit AD pathology, and AD often manifests atypically[14]. Thus, using biomarkers, such as amyloid beta (Aβ), in genetic research can help discover more precise signals and novel AD-associated loci[15].

In this study, we examined a wide range of genetic risk factors for AD using high-depth WGS data from 1559 individuals from a Korean cohort with AD. We performed a GWAS for common and rare AD-associated variants using the core AD biomarkers amyloid PET and clinical phenotypes. We compared our findings with existing Asian AD GWAS to assess the reproducibility and novelty of the identified loci. In addition, we characterized the relationships between AD and rare noncoding variants that disrupt gene regulatory elements, CNV, and rare STR expansions at the *APOE* locus. Our integrative WGS study identified the associations and cumulative effects of various genetic factors with cognitive impairment and Aβ deposition in AD.

Results

Study description

Our Korean AD WGS study recruited 1824 samples from the Korea Registries to Overcome and Accelerate Dementia (K-ROAD) project from 2017 to 2023 and generated high-depth WGS data (average of 30 × depth per sample) for these samples (Supplementary Fig. 1). After quality control, we excluded 76 samples due to low-quality data or relatedness and 189 with other types of dementia (Supplementary Fig. 2). As a result, 1559 individuals were included, comprising 655 cognitively unimpaired (CU) individuals, 590 with mild cognitive impairment (MCI), and 314 with dementia of the Alzheimer's type (DAT) (Supplementary Fig. 3, Supplementary Data S1a, S1,b). CU individuals were significantly younger than those in the MCI ($p < 2.2 \times 10^{-16}$) and DAT ($p < 2.2 \times 10^{-16}$) groups, according to the *t* tests.

From the WGS dataset, we prioritized high-quality single-nucleotide variants (SNVs) and insertions and deletions (indels), which were used for GWAS of common variants, gene-based testing of rare coding variants (Figs. 1 and 2), and genetic association testing of rare noncoding variants (Fig. 3). In addition, this dataset, was used to identify CNVs and STRs (Fig. 4).

Our dataset includes various AD biomarkers, phenotypes, and clinical diagnostic data. Aβ positivity was defined using the global centiloid values derived from Aβ PET imaging. Among the participants, 925 were classified as Aβ-positive, whereas 634 were Aβ-negative. We assessed cognitive function using the Korean Mini-Mental State Examination (K-MMSE), Clinical Dementia Rating Sum of Boxes (CDR-SB), verbal memory, and visual memory tests for the participants. The rich phenotypic data enabled various comparisons of genetic factors and a precise understanding of their effects (Supplementary Data S1c).

Identification of genes relevant to AD severity using common and rare coding variants

To perform GWAS of AD, we first performed single-variant association analysis using common variants (minor allele frequency, MAF ≥ 1%). Subsequently, gene-based association analyses were performed using rare coding variants (MAF < 1%), specifically those annotated as deleterious (Supplementary Fig. 4). Clinical diagnosis and Aβ positivity

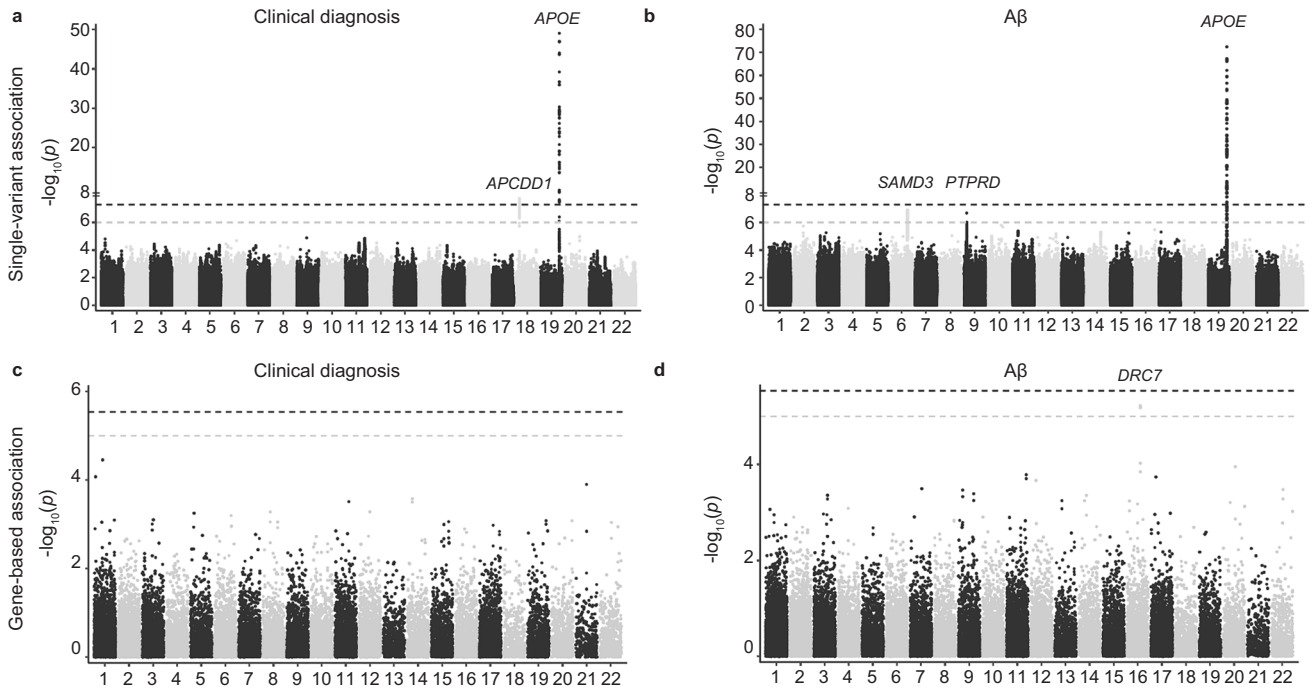

Fig. 1 | GWAS of clinical diagnosis and Aβ and the explanatory power of the identified loci. a Manhattan plot for single-variant association analysis of clinical diagnosis with a meta-analysis of Korean and Japanese cohorts. The genome-wide significance threshold is indicated by a black horizontal dashed line at *p* of 5×10^{-8}, and the suggestive threshold is indicated by a gray line at $p = 1 \times 10^{-6}$. The nearest genes were annotated for signals suggestive of significance. **b** Manhattan plot for single-variant association analysis of Aβ using the Korean WGS cohort. **c**, **d** Manhattan plots for gene-based association analysis of clinical diagnosis (**c**) and Aβ levels (**d**). The Bonferroni-corrected significance threshold is indicated by a black horizontal dashed line at $p = 2.9 \times 10^{-6}$, and the suggestive threshold is indicated by a gray line at $p = 1 \times 10^{-5}$. Genes were annotated for signals reaching suggestive significance. Two-sided Firth logistic regression was performed using REGENIE, with no correction for multiple testing. GWAS, genome-wide association analysis; Aβ, amyloid beta; WGS, whole-genome sequencing.

https://doi.org/10.1038/s41467-025-59949-y

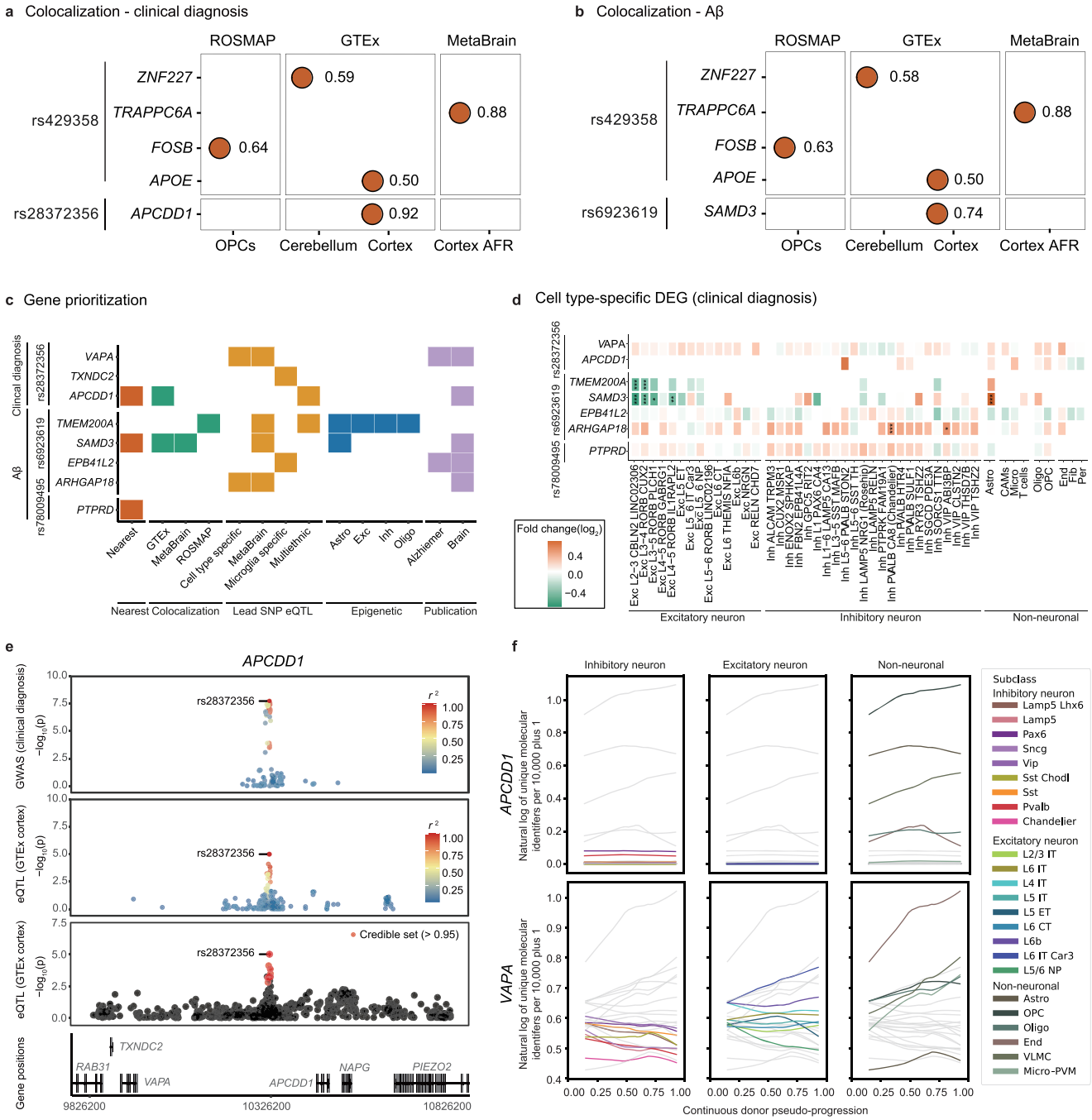

Fig. 2 | Gene prioritization and cell type-specific expression patterns of prioritized genes. a, **b** Colocalization of GWAS of clinical diagnosis (**a**) and Aβ levels (**b**) with ROSMAP, MetaBrain, and GTEx eQTLs. Colocalizations with a posterior probability above 0.5 are shown. **c** Heatmap of gene prioritization of loci reaching suggestive significance from single-variant association analysis. The heatmap includes genes nearest to lead-SNPs (red), eQTL colocalization (green), eQTL signals associated with lead-SNPs (yellow), peak-to-gene connections identified through scATAC (blue), and evidence from prior publications (purple). **d** Heatmap of DEGs derived from a previous study (Mathys et al.) according to the final cognitive consensus diagnosis across brain cell subtypes. Significance levels are indicated as $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. **e** Regional *p*lots of the clinical diagnosis GWAS and GTEx cortex eQTLs for *APCDD1* within a 500 kb window of the lead variant (rs28372356). **f** *APCDD1* and *VAPA* expression across different cell types in response to pseudo-progression of SEA-AD. Each line represents the locally weighted mean expression (LOWESS) of the supertypes with each subclass. Source data are provided as a Source Data file. Aβ, amyloid beta; ROSMAP, Religious Orders Study/Memory and Aging Project; GTEx, genotype-tissue expression; AFR, African; OPC, oligodendrocyte precursor cell; eQTL, expression quantitative trait loci; DEG, differential expressed gene; Astro, astrocyte; Exc, excitatory neuron; Inh, inhibitory neuron; CAMs, cell adhesion molecules; Micro, microglia; Oligo, oligodendrocyte; End, endothelial cells; Fib, fibroblasts; Per, pericytes; SMC, smooth muscle cell; VLMC, vascular and leptomeningeal cells; Micro-PVM, microglia and perivascular macrophages; GWAS, genome-wide association analysis; scATAC, single-cell chromatin accessibility; SEA-AD, Seattle Alzheimer's Disease Brain Cell Atlas.

phenotypes were the outcome measures. We examined the genetic association of these variants with AD using 314 individuals diagnosed with DAT, 655 individuals with CU, and 925 Aβ-positive and 634 Aβ-negative individuals.

To identify the associations between single common variants and clinical diagnoses, we increased the sample size by conducting a meta-analysis of results from other East Asian GWAS, including Japanese[16] and Korean studies. The Japanese cohort comprised 140 individuals

https://doi.org/10.1038/s41467-025-59949-y

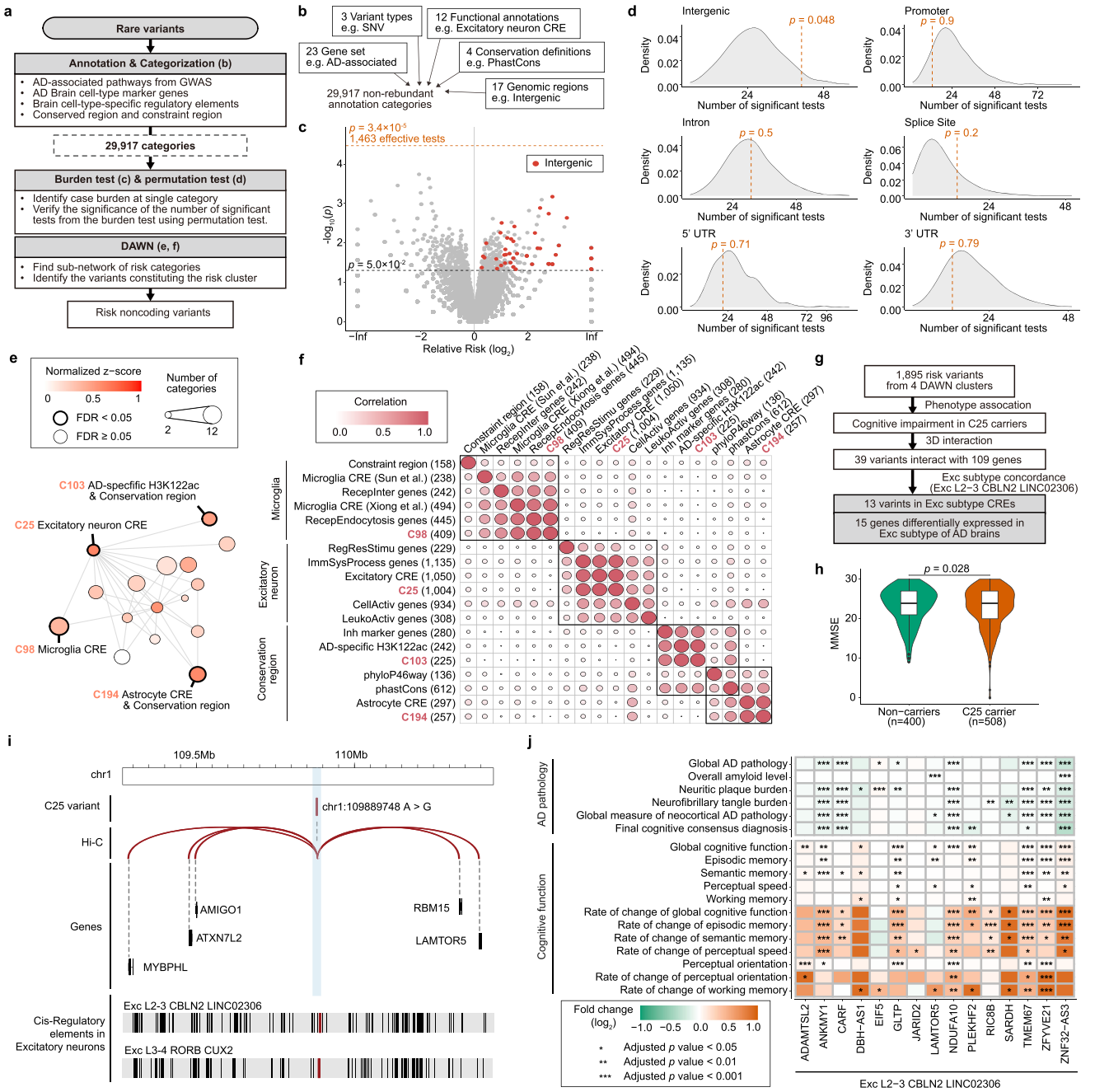

Fig. 3 | Identification of rare noncoding variants associated with AD risk. **a** Overview of rare noncoding variant analysis using CWAS. **b** Variants were annotated with 59 terms across five groups, generating 29,917 non-redundant categories. **c** Volcano plot showing burden enrichment across categories; intergenic categories (red) showed significant enrichment (RR > 1, $p < 0.05$). A Bonferroni-corrected significance threshold was applied based on 1463 effective tests. **d** Density plots show the number of significant tests for each phenotype within each noncoding category. The red lines represent nominally significant Aβ-positive-enriched categories. The permuted expected distribution is shown in gray. Significance tested by permutation ($n = 1000$). **e** Network of AD-associated intergenic category clusters. Each node represents a cluster of categories. Each node is interconnected based on correlations with disease association presented by normalized z-scores. **f** The correlation between single annotations and four risk clusters (FDR < 0.05, RR > 1 The four clusters were grouped into three terms (correlation > 0.5). Numbers of variants indicated in parentheses. **g** Schematic of variant selection within the risk cluster. **h** Comparison of K-MMSE scores between individuals carrying and not carrying the C25 variant within the Aβ-positive group. *n* represents the number of samples in each group (C25 carriers: $n = 508$; non-carriers: $n = 400$). Box plots show the median line. Box edges mark the 25th and 75th percentiles. Whiskers span 1.5 × the interquartile range. Points beyond are outliers. Two-sided linear regression was used with adjustment for sample covariates. **i** Example of a C25 variant interacting with multiple genes through Hi-C and overlapping excitatory neuron-specific regulatory elements. **j** Heatmap of genes linked to risk variants across AD phenotypes, with color gradients indicating log2-fold changes and significance marked by *. Only gene–variant pairs from excitatory neuron subtype (Exc L2–3 CBLN2 LINC02306) shown. Differential expression tested via quasi-likelihood F-test (muscat), FDR adjusted. Source data are provided as a Source Data file. AD, Alzheimer's disease; CWAS, category-wide association study; FDR, false discovery rate; RR, relative risk; C25, Cluster 25; H3K122ac, acetylate lysine 122 on H3; APP catabolism, regulation of amyloid precursor protein catabolic process (GO:1902991).

https://doi.org/10.1038/s41467-025-59949-y

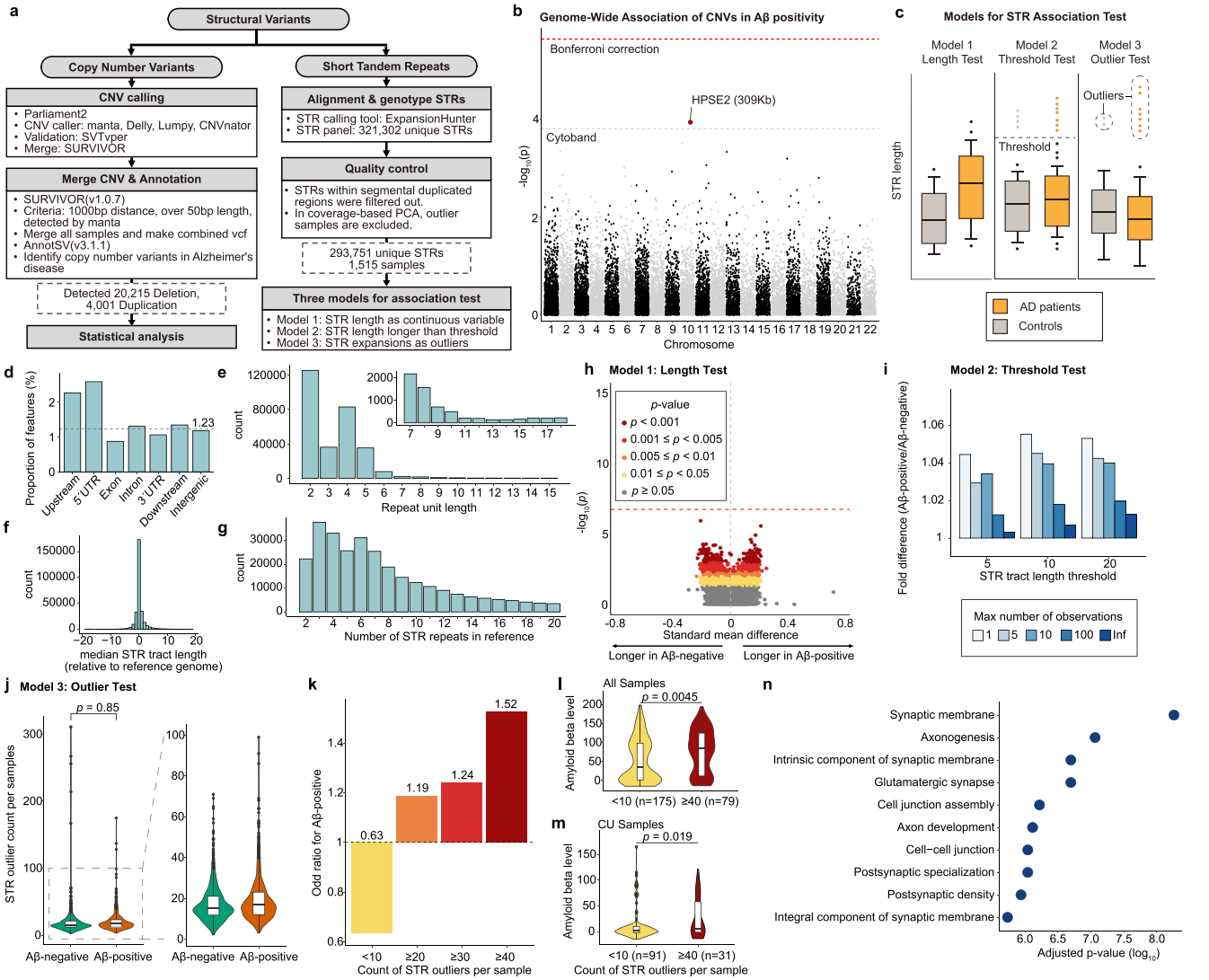

Fig. 4 | Structural variants associated with AD. a Overview of association test between structural variants and AD. **b** Manhattan plot for association analysis of CNV with Aβ positivity. Red and gray dashed lines indicate the Bonferroni ($p = 2.4 \times 10^{-6}$) and cytoband-based Bonferroni ($p = 1.6 \times 10^{-4}$) thresholds, respectively. Two-sided logistic regression was used with sample covariates. **c** Schematic of three STR association models. **d** Proportion of tandem repeats observed within each genomic region. The proportion reflects the frequency of STR presence relative to the size of the region. Dashed line: genome-wide average. **e** Histogram of STR motif length (inset: ≥ 7 bp). **f** Histogram of median STR tract lengths relative to GRCh38, genotyped by ExpansionHunter; x-axis limited to − 20 to 20. **g** Histogram of STR repeat counts in the GRCh38. **h** The association between individual STRs (x-axis: mean length difference between Aβ-positive and Aβ-negative samples divided by standard deviation). STR lengths were analyzed using two-sided logistic regression, adjusting for sample, technical, and *APOE* ε4 covariates. **i** Relative burden of STRs enriched in Aβ-positive individuals across thresholds of STR length and observation count. **j** Distribution of STR outlier counts for each sample compared across Aβ-positive ($n = 895$) and Aβ-negative samples ($n = 620$). **k** Odds ratio measuring the likelihood of Aβ positivity in individuals with different numbers of STR outliers. **l** Aβ levels of each sample were compared between groups and divided according to the threshold (< 10 or ≥ 40) for the STR outlier count. **m** Same as (**l**), but for CU samples only. **n** Gene set enrichment analysis for genes near STR outliers in samples with ≥ 40 outliers (two-sided Fisher's exact test, FDR-adjusted). The box plots in this figure show the median line. Box edges mark the 25th and 75th percentiles. Whiskers span 1.5 × the interquartile range. Points beyond are outliers. For (**j**, **l**, **m**) two-sided linear regression models were used with sample and technical covariates, and *n* refers to the number of individual samples. Source data are provided as a Source Data file. Aβ, amyloid beta; AD, Alzheimer's disease; CU, cognitively unimpaired; CNV, copy number variants; STR, short tandem repeat.

with probable AD and 798 cognitively normal older adult controls. The Korean cohort included 523 DAT individuals and 340 CU individuals, all from independent samples, and genotyping was conducted using a chip array. This meta-analysis revealed two genome-wide significant loci near *APOE* and *APCDD1* ($p = 1.81 \times 10^{-8}$) (Fig. 1a, Supplementary Fig. 5a and Table 1). For *APCDD1*, consistent signals were observed across three independent East Asian cohorts (Korean WGS, $p = 7.49 \times 10^{-5}$; Korean chip array, $p = 5.85 \times 10^{-3}$; Japanese WGS, $p = 1.79 \times 10^{-3}$). For Aβ positivity, we conducted GWAS using WGS data, identifying one significant locus, *APOE*, and two suggestive loci near *SAMD3* ($p = 1.22 \times 10^{-7}$) and *PTPRD* ($p = 2.07 \times 10^{-7}$) (Fig. 1b and Supplementary Fig. 5b). These three loci (*APCDD1*, *SAMD3*, and *PTPRD*) have not been reported in European or East Asian studies[3,17] (Supplementary Data S2a).

Rare coding variants were evaluated using gene-based association analysis. We prioritized rare coding variants, which were predicted to be loss-of-function or deleterious missense variants, and then collapsed the variants using genes for association testing. While the association analysis for clinical diagnosis did not identify any significant genes, the testing for Aβ positivity identified a suggestive gene burden for *DRC7*

Table 1 | Genetic variants significantly associated with clinical diagnosis or Aβ positivity

Gene	Locus	Type	Associated phenotype	Distance to TSS	Odd ratio (95%CI)	*p* value
APCDD1	chr18:10326201 (rs28372356)	SNV	Clinical diagnosis	−128,434	1.65 (1.39–1.96)	1.81×10^{-8}
HPSE2	chr10:98736468-98737614	DEL	Aβ	279,391–280,537	1.60 (1.25–1.99)	0.0001

Two-sided Firth logistic regression (for SNV) or logistic regression (for DEL) was performed, without correction for multiple testing.
CI confidence interval, *FDR* false discovery rate, *SNV* single-nucleotide variant, *DEL* copy number deletion, *Aβ* amyloid beta, *TSS* transcription start site.

($p = 5.99 \times 10^{-6}$) (Fig. 1c, d, Supplementary Fig. 5c, d and Supplementary Data S2b, c). Among the 25 prioritized rare coding variants in the *DRC7* gene, 18 variants were more frequently observed in Aβ-negative individuals than in Aβ-positive individuals, with 11 variants exhibiting the highest allele frequency in East Asians (Supplementary Data S2d). Two variants (chr16:57707535 and chr16:57723121) were predicted to be pathogenic using AlphaMisSense, a machine learning model that assesses missense variant pathogenicity. In addition, other computational pathogenicity predictors identified four variants by MetaSVM, 17 variants by SIFT, and 13 variants by PolyPhen. Nineteen variants showed positive PhyloP mammalian conservation scores, suggesting that most were located in conserved genomic regions.

Based on the cell type-specific differentially expressed genes (DEGs) identified in a previous study[18], *DRC7* expression was elevated in excitatory neuron subtypes with low social isolation scores and increased in astrocytes of individuals with diabetes (Supplementary Fig. 6a). *DRC7* was highly expressed in excitatory neurons and astrocytes in the Seattle Alzheimer's disease brain cell atlas (SEA-AD) (Supplementary Fig. 6b).

Prioritized genes associated with GWAS loci and their expression patterns linked with AD severity

To prioritize genes associated with the GWAS loci, we conducted statistical fine-mapping analyses through expression quantitative trait loci (eQTL) colocalization. We employed three different eQTL databases: genotype-tissue expression (GTEx) eQTL data, cell type-specific eQTL data from the Religious Orders Study/Memory and Aging Project (ROSMAP), and MetaBrain eQTL data. In the clinical diagnosis GWAS, the rs429358 locus exhibited colocalization with four genes (*ZNF227*, *TRAPPC6A*, *FOSB*, and *APOE*), and the rs28372356 locus was found to colocalize with *APCDD1* (Fig. 2a and Supplementary Data S2e). In the Aβ GWAS, the rs429358 locus was colocalized with the same four genes (*ZNF227*, *TRAPPC6A*, *FOSB*, and *APOE*), whereas the rs6923619 locus exhibited colocalization with *SAMD3* (Fig. 2b and Supplementary Data S2e). No significant colocalization was observed for rs78009495.

Next, we conducted gene prioritization for the three previously unreported identified loci (rs28372356, rs6923619, and rs78009495) (Fig. 2c and Supplementary Data S2f, i). Gene prioritization was performed based on the following five criteria: nearest gene, co-localization with eQTL, eQTL signals on lead SNPs, evidence from prior publications, and peak-to-gene connection identified with single-cell chromatin accessibility (scATAC). First, for rs28372356, which is associated with a clinical diagnosis, the nearest gene was *APCDD1*, with co-localization observed between the eQTL and GWAS signals. In addition, this lead variant exhibited significant eQTL signals for *APCDD1*, *VAPA*, and *TXNDC2*. Both *APCDD1* and *VAPA* have been reported in previous AD- and brain-related studies[19–22]. Next, for the rs6923619 variant, associated with Aβ positivity, *TMEM200A*, *EPB41L2*, *ARHGAP18*, and *SAMD3* were identified as prioritized genes (Supplementary Fig. 7). Finally, for rs78009495, associated with Aβ positivity, the *PTPRD* gene was prioritized.

We also investigated whether these genes were differentially expressed according to the clinical diagnosis using cell type-specific DEGs calculated from previous research[18] (Fig. 2d and Supplementary Data S2j). In patients with AD, *ARHGAP18* expression was found to be higher in inhibitory neurons. In addition, *TMEM200A* and *SAMD3* expression were lower in excitatory neurons, whereas *SAMD3* expression was higher in astrocytes. Among individuals carrying the rs6923619 variant, lower amyloid beta levels and *SAMD3* expression were observed in the GWAS and eQTL analyses, respectively (Supplementary Data S2f). Single-cell ATAC-seq further revealed a peak-to-gene link at the *SAMD3* locus in astrocytes, where the DEG directionality aligned with both the GWAS and eQTL findings (Supplementary Data S2h).

To further explore the association between AD and the prioritized genes, we examined the cell-type-specific expression patterns. We focused on the genes (*APCDD1*, *VAPA*, and *TXNDC2*) prioritized at the previously unreported identified significant locus (rs28372356). The rs28372356 variant exhibited significant overlapping GWAS and eQTL signals of *APCDD1* in the GTEx cortex eQTL data (PP.H4 = 0.92) and was identified as a variant in the credible causal variant set (Fig. 2e and Supplementary Data S2e). *APCDD1* was elevated in non-neuronal cell types, particularly in oligodendrocyte precursor cells (OPCs), and its expression increased with AD progression (Fig. 2f). Furthermore, *APCDD1* is predominantly expressed in non-neuronal cell types within the brain transcriptome at the single-cell level (BTS) atlas[23], with expression levels increasing with age, particularly in OPCs. (Supplementary Fig. 8). Similarly, *VAPA* was predominantly expressed in non-neuronal cell types, especially astrocytes, and its expression increased with AD progression (Fig. 2f). In contrast, *TXNDC2* exhibited a consistently low expression in all brain cell types (Supplementary Fig. 9).

Identification of rare noncoding variants in excitatory neuron regulatory elements associated with AD severity

Beyond the GWAS of common and rare coding variants, we leveraged the WGS data to investigate rare noncoding AD-associated variants. We performed a category-wide association study (CWAS)[24,25], a statistical framework integrating multiple functional annotations relevant to disease pathology (a priori hypothesis) and combining multiple categories using these annotations. CWAS can perform category-based collapse analysis for noncoding associations with appropriate multiple comparisons. For CWAS analysis, we utilized 19,266,739 rare variants (MAF < 0.1%) from 1559 samples. We categorized rare variants using AD-relevant annotations, such as cis-regulatory elements (CREs) from postmortem AD brain samples[26,27], biological pathways enriched for known GWAS loci[3] (e.g., TNF-mediated signaling, endocytosis, and immune activation), and conserved sequences (Fig. 3a and Supplementary Data S3a). For Aβ positivity analysis, we categorized rare variants from Aβ-positive and Aβ-negative samples into 29,917 categories, each with at least one variant (Fig. 3b and Supplementary Data S3b).

We compared the burden across each of the 29,917 categories. Although no category showed enrichment after multiple testing corrections (Supplementary Data S3b), 287 categories showed nominal significance ($p < 0.05$, relative risk (RR) > 1, binomial test), and intergenic categories accounted for the highest numbers, with 45 out of 287 categories (Fig. 3c). Permutation tests confirmed statistical significance in the intergenic categories ($p = 0.048$), but not in other noncoding regions (Fig. 3d and Supplementary Data S3c). Applying the same analysis to clinical diagnosis did not reveal significant signals in the noncoding regions. Power estimation indicated limited power for clinical diagnosis, directing subsequent analyses toward Aβ positivity-associated categories (Supplementary Fig. 10).

We constructed a network based on the correlation of sample counts across 2,074 intergenic categories to elucidate the functional

annotations associated with the intergenic categories. Consequently, we found 194 clusters of these categories, including four clusters associated with Aβ positivity (false discovery rate, FDR < 0.05; RR > 1) (Fig. 3e and Supplementary Data S3d). After analyzing the correlations between clusters and annotation terms, we identified functionally distinct groups of intergenic variants (Fig. 3f). The most significant was Cluster 25 (FDR = 8.7×10^{-5}; RR = 1.1), containing rare noncoding variants within excitatory neuron-specific CREs identified in postmortem AD brain samples[26] (Supplementary Data S3e).

We assessed four AD clinical phenotypes for each cluster to investigate the association between the variants in the four risk clusters and AD (Fig. 3g and Supplementary Fig. 11). Although none of the 16 tests passed FDR, we prioritized Cluster 25 (C25) for its strongest nominal significance and potential association with increased AD severity. Carriers of Aβ-positive-associated variants in C25 exhibited decreased K-MMSE scores, compared with other Aβ-positive samples ($p = 0.028$, Fig. 3h). C25 comprises variants located in excitatory neuron CREs associated with cognitive function in AD in a previous study[18]. We examined the interactions between each variant and gene(s) using Hi-C data to identify the genes affected by these variants[28]. We identified that among the 63 Aβ-positive-only variants in C25, 39 variants interacted with 109 genes. To further elucidate the impact of each variant, we identified C25 variants within specific excitatory CREs[26] (Fig. 3i) and examined the association between gene expression and phenotypic characteristics within the same excitatory subtypes using available AD single-cell transcriptome data[18]. Alterations in the expression of the 15 genes linked to the 13 variants were associated with cognitive impairment in the excitatory neuron subtype, Exc L2–3 CBLN2 LINC02306 (Fig. 3j and Supplementary Data S3f).

Comprehensive SV analyses in AD reveals rare STR expansion burdens linked to Aβ

To investigate the association between SVs and AD risk, we conducted association tests for CNVs and STRs. We identified 24,516 CNVs in 1555 participants, including 20,515 deletions and 4001 duplications (Fig. 4a and Supplementary Fig. 12a, b). Logistic regression was employed to assess the association of CNVs with Aβ positivity and clinical diagnosis (Fig. 4b, Supplementary Fig. 12c and Supplementary Data S4a). While no significant associations were observed at the Bonferroni-corrected significance level, the deletion encompassing *HPSE2* was associated with Aβ positivity after cytoband-based Bonferroni correction ($p = 0.036$). We visually confirmed the presence of deletions in the *HPSE2* region (Supplementary Fig. 12d). Although the deletion encompassing *ZNF7* was significantly associated with clinical diagnosis after cytoband-based Bonferroni correction ($p = 0.031$), this was deemed a false positive because of false-positive deletion calls observed through visualization (Supplementary Fig. 12e).

Next, we identified 293,751 distinct STRs across 1515 samples (Fig. 4a). We assessed their association with AD using three models (Fig. 4c). Model 1 compared STR lengths between cases and controls, Model 2 examined the burden of STRs exceeding a specific length threshold, and Model 3 identified STR outliers by evaluating length differences within our cohort and analyzing their count differences between cases and controls. STRs were more prevalent within 1000 bp upstream of protein-coding genes and in the 5′ UTR compared to the genome-wide average, with proportions of 2.27% and 2.59%, respectively (Fig. 4d). The repeat unit lengths of the STRs were predominantly within 2–5 base pairs (Fig. 4e). Examination of the median length of each STR revealed that 59% aligned closely with the reference genome with no length differences (Fig. 4f). In the reference genome, STRs of ≤ 20 base pairs explained 96% of the distribution (Fig. 4g). No differences were observed in sample mean coverage in Aβ-positivity and clinical diagnosis (Supplementary Fig. 13).

Using Model 1 (Length Test) for Aβ positivity, no significant STRs were identified (Fig. 4h and Supplementary Data S4b). However, when *APOE* ε4 status was not considered, one STR was significant, consistent with findings from a previous European study[13] (chr19:44921097-44921125-TTTA; $p = 7.48 \times 10^{-39}$; Supplementary Fig. 14). This STR was in LD with the *APOE* ε4 allele ($r^2 = 0.56$). Model 2 (Threshold Test) for Aβ positivity found no significant STRs, excluding the STR (chr19:44921097-44921125-TTTA) identified in Model 1 (Supplementary Data S4c). Additionally, we investigated the association between STR expansion frequency and Aβ positivity. We observed that for each threshold (≥ 5, ≥ 10, ≥ 20), rare STRs tended to have more burden in Aβ positivity (Fig. 4i). Model 1 for clinical diagnosis showed no significant associations, and Model 2 for clinical diagnosis exhibited no trend of burden for rare STRs; the analysis based on Aβ positivity showed greater concordance with results from the previous European study[13] than that for clinical diagnosis. Analyses of the two models indicated that individual STRs are less likely to drive AD risk, whereas the burden of rare STRs may increase the risk for Aβ positivity.

Finally, to investigate whether rare STR expansions were associated with AD, outliers of each STR were called (Fig. 4c, Model 3: Outlier Test). Of 293,751 STRs, 15,910 were identified as outliers. No significant differences were found when comparing the number of STR outliers in Aβ positivity (Fig. 4j). The average outlier count was slightly higher in Aβ-positive samples (Aβ-positive = 19.59, Aβ-negative = 19.36, Supplementary Data S4d). The odds ratio for Aβ positivity increased with the count threshold and was independent of the *APOE* ε4 ratio (Fig. 4k) ($p = 1.92 \times 10^{-7}$, Supplementary Fig. 15). To validate these findings, we compared samples with < 10 and those with ≥ 40 outliers to assess differences in Aβ levels. Samples with ≥ 40 outliers had significantly higher Aβ levels ($p = 0.0045$, Fig. 4l). Moreover, in samples diagnosed as CU, those with ≥ 40 outliers had significantly higher Aβ levels ($p = 0.019$, Fig. 4m). Only one outlier STR was found in LD with primary risk SNPs, and its exclusion did not alter the results, confirming that these STRs were independent risk factors (Supplementary Data S4e). Pathway enrichment analysis showed that the genes for STR outliers were enriched for synaptic functions, such as synaptic membranes (GO:0097060) (Fig. 4n and Supplementary Data S4f).

Phenotypic association of genetic factors in AD

We assessed the increase in explanatory power of the variants identified in this study. We compared the phenotypic variance explained by previously reported variants in European studies[3] with that explained by the variants identified in our study. We calculated incremental r^2 to measure the increase in r^2 when these variants were added. For clinical diagnosis, the incremental r^2 was 12.8% with only the European loci (Fig. 5a). This increased to 14.7% upon including loci identified in the Korean cohort and further increased to 16.6% with the inclusion of RVs and SVs. For Aβ, the incremental r^2 increased from 7.5% with European loci to 9.8% with the addition of Korean loci and reached 10.7% after the inclusion of RVs and SVs.

To understand the genetic complexity of AD, we explored the phenotypic association of diverse genetic factors across *APOE* ε4 genotype, polygenic, rare noncoding variants, and SVs. We excluded rare coding variant carriers because these variants exhibited a protective effect opposite to that observed for other variants. Individuals with single variants among the rare noncoding variants of C25 were classified as rare noncoding variant carriers ($n = 862$). SV carriers were characterized as individuals with either CNVs within the *HPSE2* gene region ($n = 10$) or STRs expanded by ≥ 40 repeats ($n = 82$).

To examine polygenic effects, we employed the effect sizes of variants as weights in the PRS from large-scale European GWAS data, as previous studies have demonstrated transferability from European to East Asian populations[29]. We calculated polygenic risk scores (PRSs) using European GWAS of clinical diagnosis[3] ($n = 487{,}511$) and Aβ positivity[15] ($n = 11{,}816$). Individuals with CU had significantly lower PRSs derived from the clinical diagnosis GWAS than those with DAT ($p = 0.0005$) or MCI ($p = 0.0334$) (Supplementary Fig. 16a). Clinical

https://doi.org/10.1038/s41467-025-59949-y

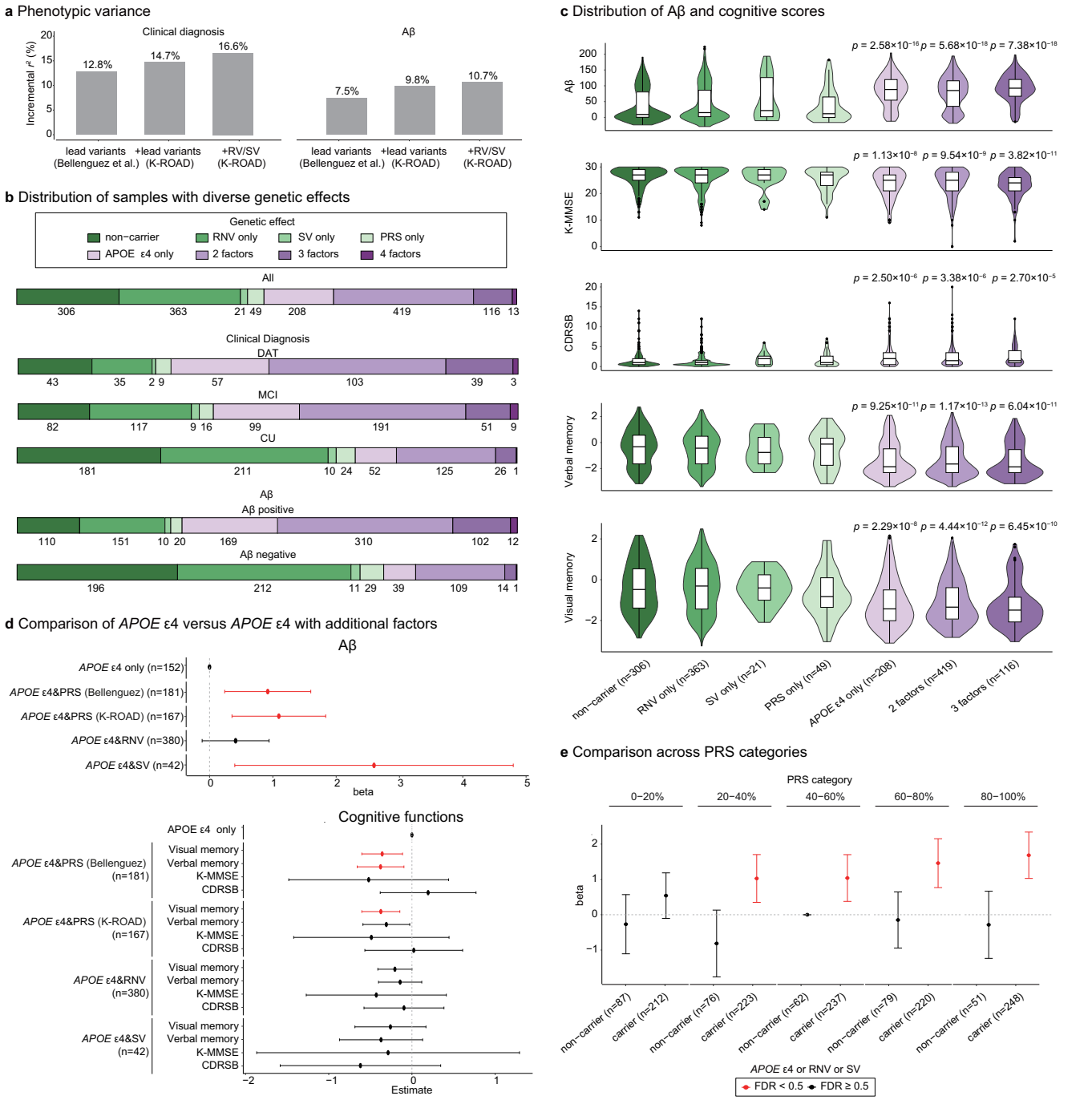

Fig. 5 | Diverse effects of genetic variants on Aβ levels and cognitive function. **a** Bar plots illustrating the incremental r^2 for clinical diagnosis and Aβ in comparison to previously reported variants (Bellenguez et al.), previously unreported identified variants, RVs, and SVs (K-ROAD). **b** Stacked bar plot depicting the distribution of samples categorized as *APOE* ε4 carriers, PRS top 20% group, RNV, or SV carriers based on clinical diagnosis status and Aβ positivity. Cases possessing two or more genetic factors were categorized into groups according to the number of factors present: two, three, or four factors. **c** Violin plots illustrating the distribution of Aβ, K-MMSE, CDR-SB, verbal memory, and visual memory scores in each group. Box plots indicate the median (center line) and interquartile range (bounds of the box). **d** Forest plot comparing the differences in Aβ positivity and cognitive function markers between groups carrying only *APOE* ε4 ($n = 437$) and those with high PRS, RNV, or SV along with *APOE* ε4. PRS calculations were based on variants identified in PRS calculations were based on variants identified in a prior European GWAS (Bellenguez et al.) or variants from the current study (K-ROAD). **e** Forest plot comparing differences in Aβ positivity within each PRS quintile for carriers and non-carriers of *APOE* ε4, RNV, and SV. Two-sided logistic or linear regression was performed to assess statistical significance in the violin and forest plots, adjusting for age, sex, batch, education, and the top 10 principal components of genetic ancestry, followed by FDR correction for multiple comparisons. FDR-adjusted P-values are presented. Error bars represent 95% confidence intervals. n indicates the number of independent participants. Source data are provided as a Source Data file. Aβ, Amyloid beta; K-ROAD, Korea Registries to Overcome and Accelerate Dementia Research; RV, rare variants; SV, structural variants; RNV, rare noncoding variant; PRS, polygenic risk score; DAT, dementia of Alzheimer's type; MCI, mild cognitive impairment; CU, cognitively unimpaired; K-MMSE, Korean Mini-Mental State Examination; CDRSB, Clinical Dementia Rating Sum of Boxes.

diagnosis ($p = 0.0005$) and Aβ positivity ($p = 0.0282$) were significantly correlated with PRS (Supplementary Fig. 16b). Individuals with higher PRS had decreased cognitive functioning based on K-MMSE ($p = 1.47 \times 10^{-3}$), visual memory ($p = 2.40 \times 10^{-3}$), verbal memory ($p = 2.03 \times 10^{-5}$), and CDR-SB ($p = 8.40 \times 10^{-4}$) scores (Supplementary Fig. 16c). Aβ-negative samples had lower PRSs—calculated from GWAS based on Aβ positivity—than Aβ-positive samples ($p = 2.49 \times 10^{-8}$) (Supplementary Fig. 16d). Individuals with high Aβ PRS were at higher risk for clinical diagnosis ($p = 5.65 \times 10^{-8}$) and Aβ positivity ($p = 2.49 \times 10^{-8}$), with reduced cognitive functioning based on K-MMSE ($p = 1.06 \times 10^{-5}$), visual memory ($p = 3.67 \times 10^{-7}$), verbal memory ($p = 7.86 \times 10^{-9}$), and CDR-SB ($p = 3.36 \times 10^{-5}$) scores (Supplementary Fig. 16e, f). In subsequent analyses, we used the PRSs calculated from the Aβ GWAS, as they exhibited lower P-values. To set PRS thresholds, we divided samples into percentiles of 5–40% and compared the risks for AD or Aβ positivity between the high PRS group and the remaining samples. We observed the highest odds ratio when contrasting the top 20% of PRS samples with the others and therefore set the threshold at 20% in all subsequent analyses (Supplementary Fig. 16g, h). We then examined the distribution of individuals with each genetic factor based on clinical diagnosis, Aβ positivity, or cognitive functioning (Fig. 5b and Supplementary Figs. 17, 18a). In the CU, Aβ-negative, and high cognitive function groups, the proportion of non-carriers was high. In contrast, in the DAT, Aβ-positive, and low cognitive function groups, *APOE* ε4 carriers and individuals with two or more genetic factors were more prevalent. When comparing with non-carriers in each group, significant associations were observed across all trait markers with *APOE* ε4 only, two factors, and three factors (Fig. 5c).

Next, we investigated whether the presence of other genetic factors besides *APOE* ε4 affected Aβ levels or cognitive functioning compared to carrying *APOE* ε4 alone. For this analysis, the PRS was calculated using the three lead SNPs (rs28372356, rs6923619, and rs78009495) identified in this study (K-ROAD). Compared with the group with only *APOE* ε4, the group with high PRS or SV along with *APOE* ε4 exhibited high levels of Aβ (Fig. 5d, Supplementary Fig. 18b, c). In addition, having high PRS alongside *APOE* ε4 was associated with lower levels of visual memory and verbal memory function.

Lastly, we stratified PRS into quintiles and compared the odds ratios for Aβ positivity between individuals with and without *APOE* ε4, rare noncoding variants, or SVs (Fig. 5e). Carriers of *APOE* ε4, rare noncoding variants, or SVs exhibited significantly higher odds ratios in the 20–40, 40–60, 60–80, and 80–100% PRS quintiles compared with the reference group, which was the non-carrier group in the 40–60% PRS quintile. In contrast, non-carriers did not show significant differences in Aβ positivity across PRS groups relative to the reference group.

Discussion

We performed a comprehensive WGS analysis to identify the various genetic factors involved in AD and their contributions to the AD phenotype. We identified three previously unreported loci with significant or suggestive genome-wide associations. Along with common variations, we examined rare noncoding variations and identified significant associations with underlying excitatory neuron-specific CREs. Integrating various AD biomarkers and phenotypic measures helped identify cognitive function or Aβ levels as core domains, where various genetic risks from common to rare variants were involved. Subsequently, a comparison of the phenotypic relationship across PRS, rare coding, noncoding, and SV variants revealed a complete landscape of genotype-phenotype associations underlying AD in a Korean cohort.

Using a large-scale East Asian WGS dataset, we identified signals near the *APCDD1*, *SAMD3*, and *PTPRD* genes that were not reported in European GWAS[3]. Postmortem brain samples from patients with AD showed reduced *SAMD* and *TMEM200A* expression in excitatory neurons associated with a higher risk of AD and lower resilience to cognitive decline, suggesting that the loss of function of these genes may underlie AD pathogenesis. We identified rare coding variants of the *DRC7* gene associated with Aβ positivity. *DRC7* expression is elevated in excitatory neuron subtypes of patients with low levels of social isolation. Although *DRC7* is not associated with dementia or AD, it is involved in neurodegenerative diseases[30]. Dynein dysfunction may disrupt Aβ clearance[31]. Rare variants of *DRC7* may interfere with dynein function.

Our findings suggest a cumulative effects model wherein genetic factors jointly affect susceptibility for AD or Aβ accumulation. Among *APOE* ε4 carriers, the increased polygenic burden may affect phenotypic severity in cognitive functioning or Aβ levels. Similar patterns were observed in *APOE* ε4 carriers with SVs in Aβ levels. However, variations in the individual factors did not cause significant phenotypic changes, indicating a cumulative effect on AD development.

Furthermore, phenotyping using Aβ PET imaging revealed signals that were distinct from those identified using clinical diagnosis. The signal from the *SAMD3* region, identified in the GWAS with the Aβ phenotype, possessed more biological implications according to post-GWAS analysis. Moreover, when PRS was calculated using previous GWAS results, despite the Aβ GWAS having approximately 40 times fewer samples than the clinical diagnosis GWAS, both PRSs effectively captured all AD-related phenotypes. In addition, our STR analysis showed a robust association for Aβ positivity, including the same STR locus identified in a recent European WGS study[13].

Despite these successes, our study has several limitations. First, the eQTL and single-cell transcriptome data used in this study were derived primarily from European populations. Future studies should produce large-scale gene expression data from diverse populations and investigate the consequences of genetic variations identified in non-European samples on gene expression regulation. Second, genetic factors jointly affect AD risk liability; however, it is unclear whether such genetic factors disrupt the same AD-related pathway. Previous large-scale GWAS reported common variations in APP metabolism, microglia, and immune activation pathways[3,4]. Genes regulated by rare noncoding variants were specific to excitatory neurons, with GWAS loci and STR expansions also showing enrichment for excitatory neurons and synaptic pathways. Although this might indicate the nature of genetic heterogeneity underlying AD, further studies with larger sample sizes should address the functional and molecular convergence of AD risk factors. CWAS analysis identified a potential association between excitatory neuron CRE variants and AD. However, single-category analyses lacked statistical significance, likely due to the complexity of noncoding variants and the limited sample size. To address this issue, DAWN analysis was conducted across multiple categories, yielding significant results. Nevertheless, the small sample size remains a key challenge, reducing statistical power and increasing the risk of false positives and negatives[25]. Future studies should incorporate larger cohorts[32]. In addition, noncoding regulatory variants exhibit context-dependent effects, thus meriting functional validation, such as through CRISPR-based perturbation[33], to confirm their biological significance. In addition, Cytoband-based Bonferroni multiple corrections were applied to the CNV analysis. We also visually verified the presence of CNVs in the identified regions using samplot and confirmed the absence of false positives in the QQ plot. However, due to the limited statistical power of the CNV analysis, these results, such as the *HPSE* deletion, should be cautiously interpreted, and further validation in larger datasets or independent cohorts is necessary to confirm their significance.

In summary, this comprehensive WGS study of a Korean AD cohort identified various genetic variations associated with AD and its core biomarker, cerebral Aβ deposition. Our findings suggest the cumulative effects of such genetic variations on AD pathology and support the need for future studies in diverse ancestral populations.

Methods

Ethics

This study complied with all relevant ethical regulations for research involving human participants and was conducted in accordance with the criteria set by the Declaration of Helsinki. This study received approval from the institutional review board (IRB) of Samsung Medical Center, and written informed consent was provided by all participants, and no financial or material compensation was offered for participation. WGS and genotyping using microarray were conducted using blood samples obtained from the participants. The collection, storage, and analyses of biospecimens, genetic data, and data as part of the K-ROAD were approved under the Samsung Medical Center; IRB No. 2022-07-092. All data were handled in accordance with relevant data protection and privacy regulations.

Study population

A total of 1824 individuals of Korean descent with available WGS data were recruite from a Korean dementia hospital-based cohort (K-ROAD). As an open cohort with ongoing data accumulation, the K-ROAD aims to develop a genotype-phenotype cohort to accelerate the development of novel diagnostic and therapeutic techniques for AD and other related dementias. Overall, 25 university-affiliated hospitals in South Korea participated in the K-ROAD cohort. Eligible participants were individuals with a spectrum of Alzheimer's clinical syndrome − CU, MCI, or DAT − who underwent amyloid PET imaging. This study was approved by the institutional review board, and written informed consent was obtained from all participants. WGS was conducted using blood samples obtained from participants.

Phenotype definitions

All the participants underwent clinical interviews, neurological examinations, neuropsychological testing, and brain magnetic resonance imaging (MRI). After these evaluations, clinical diagnoses were established by consensus among multidisciplinary teams. CU participants were selected based on the following criteria: (1) absence of medical history that is likely to affect cognitive function based on Christensen's health screening criteria[34] and (2) absence of objective cognitive impairment observed on any cognitive domain (above the −1·0 standard deviation (SD) of age- and education-matched norms in memory and above −1·5 SD in other cognitive domains)[35]. Participants with MCI met the specified criteria[36]: (1) subjective cognitive complaints by the participants or caregiver; (2) objective cognitive impairment in any cognitive domain (below the −1·0 SD of age- and education-matched norms in memory and/or below −1·5 SD in other cognitive domains); (3) no significant impairment in activities of daily living; and (4) no dementia. The participants diagnosed with DAT fulfilled the NIA-AA diagnostic criteria[37]. All participants underwent clinical interviews, neurological examinations, neuropsychological testing, and brain MRI. After these evaluations, clinical diagnoses were determined through agreement among the multidisciplinary teams.

All participants underwent Aβ PET with either 18F-florbetaben (FBB) or 18F-flutemetamol (FMM). To quantify the Aβ burden on PET scans as centiloids (CL), we followed the method described by Klunk et al.[38]. All imaging analyses for the K-ROAD study were conducted at the Samsung Medical Center laboratory, which served as the core center. Aβ positivity was defined using a threshold of 40 on the global centiloid scale derived from Aβ PET imaging. In cases where global centiloid values for individuals were unavailable, an expert visual assessment was used to determine Aβ positivity.

To determine global cognitive function, we used the K-MMSE and CDR-SB. We also used Seoul Verbal Learning Test (delayed recall) and Rey Complex Figure Test (RCFT) (delayed recall) scores as representative measures of verbal and visual memory functions, respectively.

WGS data alignment and variant calling

Genomic DNA was extracted from blood samples using the QIAmp DNA Mini Kit (QIAGEN). For sequencing, library preparation was performed with a TruSeq® DNA PCR-Free Library Prep Kit (Illumina), and DNA size selection was performed via Covaris ultrasonication using 1 μg of input DNA for an average insert size of 350 bp. Sequencing was performed at an average depth of 30 × with paired-end sequencing using a NovaSeq 6000 instrument with an S4 flow cell.

The paired-end raw sequencing data were initially processed via quality trimming, adapter trimming, removal of short sequences, and hard trimming using Trim Galore software (RRID:SCR_011847) (https://github.com/FelixKrueger/TrimGalore). Subsequently, the sequenced reads were aligned to the hg38 reference genome using the BWA-MEM software[39,40]. After alignment, duplicates were removed using the GATK (v.4.2.4.1) MarkDuplicate[41–43]. Base quality score recalibration was conducted using BaseRecalibrator with a WGS interval contig, indels from Mills and 1000 G gold standard, known indels from the *Homo sapiens* assembly38, and high-confidence SNPs from the 1000 G phase1. Germline SNP and indel calling were performed using HaplotypeCaller, and base quality score recalibration was conducted using ApplyBQSR. Genotype GVCFs were used to produce gVCFs, and a DB folder was created using Genomics dbimport. Finally, the gVCFs from the DB folder were combined using CombineGVCFs.

WGS quality control

The quality control (QC) of variants and samples was performed using Hail (v0.2.68) (https://github.com/hail-is/hail), except for principal component analysis (PCA) and relatedness analysis, which were performed using PLINK v1.90[44] and KING 2.0[45], respectively. First, we performed pre-filtering and genotyping QC. Prefiltering included splitting multi-allelic variants, variant quality score recalibration (VQSR) filtering, including allele counts greater than 0, removal of LCR regions, and removal of a previously defined duplicated sample (n = 1). Genotype QC was performed using the following criteria: genotype quality (GQ) (GQ ≥ 20), allele balance (AB) (hetero-variants AB ≥ 0.2 and ≤ 0.8, homo variants AB ≥ 0.9), and read depth (DP) (autosomal DP ≥ 10 and ≤ 200, chrX DP (female) ≥ 10 and ≤ 200, chrX DP (male) ≥ 5 and ≤ 200, chrY DP ≥ 5 and ≤ 200).

For sample QC, we used high-confidence variants based on the following criteria: biallelic variants, high call rates (> 0.95), and common single SNVs (allele frequency > 0.1%). We excluded samples with low coverage (mean depth ≥ 15) and low sample-level call rates (missingness ≥ 0.9). Samples with unmatched sex (f stat for females < 0.2, f stat for males > 0.8) or ambiguous sex (fstat > 0.3 and < 0.8) were excluded. We then applied different variant QC criteria and included only autosomal and biallelic variants with high call rates (> 0.95) and allele frequencies (> 5%). Relatedness was calculated using KING, and samples up to the second degree were excluded, keeping only one sample. The sample was removed as follows. First, batch 3 samples were excluded because they were excluded from the STR analysis. Samples diagnosed with AD were prioritized, followed by those of older age. Finally, we prioritized the inclusion of samples in sequencing batches 1, 2, and 4, as they were sequenced earlier. After removing related samples, PCA was performed using PLINK v1.90[44]. Next, non-reference genotype concordance was calculated for samples with available chip array data ($n = 947$), and samples with concordance below 0.5 were excluded.

After sample QC, samples that failed the sample QC were removed from the raw VCF file. We repeated the prefiltering and genotype QC, including the VQSR, LCR region, allele balance, GQ, and DP. We excluded variants with excess heterozygosity (inbreeding coefficient < − 0.3), high missing rates (call rate <0.9), and high Hardy-Weinberg equilibrium (HWE) with control samples (HWE > 1e-09). We divided the variants into SNPs and indels, and QC procedures were conducted separately. SNPs were filtered based on QD ≥ 2, SOR ≤ 3, FS ≤ 60,

MQ ≥ 50, MQRankSum ≥ − 12.5, and ReadPosRankSum ≥ − 8.0. Indels were filtered based on QD ≥ 2, FS ≤ 200, ReadPosRankSum ≥ − 20, MQ ≥ 50, and MQRankSum ≥ − 12.5.

Finally, after merging the SNPs and indels, we conducted QC of the final sample. Samples exceeding five SDs of the mean in any criterion, such as the number of SNPs, insertions, deletions, transition/transversion (Ti/Tv), hetero/homo variants, ratio of insertion-deletion, hetero-homo variants, and Ti/Tv, were excluded.

Variants were annotated using the Variant Effect Predictor (VEP)[46] v108. The pLoF variants were annotated using the LOFTEE plugin[47], and the dbSNP154, REVEL, metaSVM, and CADD Phred scores were imported using the dbNSFP (version 4.3) plugin[48,49]. The SpliceAI scores were annotated using the SpliceAI plugin[50].

Chip array data quality control

Genomic DNA was genotyped using an Asian Screening Array Chip (Illumina). Quality control procedures were as previously described[29]. Briefly, we performed variant QC based on criteria including MAF and HWE. We conducted sample QC, including the removal of duplicates, related samples, and PCA outliers. Imputation was performed using the Korean Imputation Service of the CODA. Samples identified as duplicates or those related to individuals in the WGS data were excluded. After excluding samples with missing phenotypes or other types of dementia, 1560 samples remained. Of these, 697 individuals with MCI were further excluded, resulting in 340 individuals with CU and 523 individuals with DAT for the association analysis.

Single-variant- and gene-based association analyses

Single-variant association analyses were conducted using REGENIE[51] (v2.2.4) with the leave-one-out cross-validation (LOOCV) method. Variants with a minor allele count (MAC) > 31 were included. The association analyses were performed for clinical diagnosis and Aβ positivity, with age, sex, sequencing batch, and principal components (PC) 1 to 10 as covariates. For the clinical diagnosis, we performed a meta-analysis of an independent Korean cohort (n = 863) using a chip array and a Japanese cohort (n = 938) from a previous study[16]. The samples from individuals in the Korean WGS cohort were excluded. We integrated the WGS data with the chip array data using PLINK v1.90[44], calculated relatedness using KING, removed related samples up to the second degree, and conducted a GWAS using REGENIE. For the Japanese cohort, variants with a MAC > 90 were used. We performed a meta-analysis using METAL[52], with STDERR as a single genomic control and heterogeneity option. Variants that were present in all datasets were included in the meta-analysis (n = 3,201,142).

Gene-based analyses were performed for clinical diagnosis and Aβ positivity using REGENIE with age, sex, sequencing batch, and PC 1 to 10 as covariates. We included rare variants with an alternative allele frequency (AAF) of 1% or lower. Annotation using VEP was employed to mask the gene-based analysis. Variants were defined based on their REVEL score (high ≥ 0.75, low ≥ 0.5), SpliceAI (high ≥ 0.8, low ≥ 0.5), and CADD score (deleterious ≥ 20). The masks are defined as follows:

Mask1 = pLoF (LOFTEE)
Mask2 = Mask1 + Missense variants (MetaSVM Deleterious and REVEL High)
Mask3 = Mask1 + Missense variants (MetaSVM Deleterious or REVEL High)
Mask4 = Mask1 + Missense variants (MetaSVM Deleterious/Tolerate or REVEL High/Low)
Mask5 = Mask4 + Splicing variants (SpliceAI High)
Mask6 = Mask4 + Splicing variants (SpliceAI High/Low)
Mask7 = Mask6 + Deleterious variants (CADD Deleterious)

Manhattan plots were generated using the ggmanh package (version 1.6.0) in R software (version 4.3.1)[53]. Regional plots for the GWAS and eQTL summary statistics were generated using the cowplot[54] package, with LD calculated using LDlinkR[55] based on the European LD reference panel. In addition, the gene regulatory regions were plotted using the UCSC genome browser with annotations[56] from GENCODE v44, GeneCard, ENCODE[57], and GeneHancer. The distance to the transcript start site was calculated based on representative transcripts from RefSeq and GENCODE in the UCSC Genome Browser. PhyloP conservation, AlphaMisSense, MetaSVM, SIFT, PolyPhen, SpliceAI, and population-specific allele frequencies from the gnomAD database were annotated using VEP.

Gene prioritization and statistical colocalization analysis

Gene prioritization. Gene prioritization was conducted based on the following four criteria: nearest genes, eQTL, publications, and scATACs. First, the genes nearest to the lead variant were identified using the UCSC genome browser. Second, we identified genes with significant eQTL signals and lead variants across four datasets, including cell-type specific[58], MiGA[59], multi-ethnic[60], and MetaBrain[61] eQTLs. Thirdly, publication-based prioritization was established based on whether the gene was mentioned with 'AD' or 'brain' in PubMed searches as of March 25, 2024. Finally, we defined the prioritized genes based on peak-to-gene associations previously calculated using scATAC-seq in the ROSMAP cohort[26]. When the lead variant and variants in high LD with the lead variant (LD r^2 > 0.5) were linked to a gene through peak-to-gene connections, the gene was prioritized. For gene prioritization, only consensus coding sequence genes were considered; pseudogenes and noncoding genes were excluded. Gene prioritization was visualized using the circlize package[62] (version 0.4.16) in R software (version 4.3.1).

Differential expression analysis. To determine whether the prioritized genes were differentially expressed in association with clinical diagnosis or related phenotypes, cell type-specific DEGs described by Mathys et al.[18] from the ROSMAP cohort were used. DecontX- and RUVr-adjusted DEGs were obtained from the AD/Aging Brain Atlas (https://compbio.mit.edu/ad_aging_brain/). Heatmap visualization was performed using the ggplot2 package (version 3.5.0) in R (version 4.3.1)[63].

Cell type-specific expression and gene trajectories. Data and images of cell type-specific expression and gene trajectories were obtained from SEA-AD provided by the Allen Brain Map[64]. The gene expression trajectory viewer available through the SEA-AD web application was used (https://sea-ad.shinyapps.io/ad_gene_trajectories/). Briefly, gene expression was measured using scRNA-seq in approximately 1.7 million cells from the medial temporal gyrus (MTG). Pseudo-progression scores were computed using machine learning methods based on the quantitative staining of key pathological proteins in the MTG.

Developmental trajectory of the human brain. Developmental trajectories were analyzed using the BTS database[23]. Briefly, data from 114 human postmortem brain samples spanning the early fetal stages to late adulthood were integrated to investigate cell-type-specific gene expression across developmental stages.

Single-cell RNA sequencing analysis of Korean AD postmortem samples

Single-cell transcriptomic data were generated using dorsolateral prefrontal cortex samples from 15 individuals who underwent autopsies at the Samsung Medical Center. Of the individuals, 9 were Aβ-positive and 6 were Aβ-negative based on the Consortium to Establish a Registry for Alzheimer's disease (CERAD) neuropathological criteria (none to sparse: negative, moderate to frequent: positive). To obtain the gene count, we used the Cell Ranger software[65] (v.6.1.2) (10 × Genomics) with the GRCh38 genome. The Cell Ranger count pipeline (including pre-mRNA) was used to account for the unspliced nuclear

transcripts. The gene count matrix of all libraries was generated using the Cell Ranger Aggr pipeline with default parameters in Cell Ranger 3.0 to call cell barcodes.

Single-cell RNA sequencing, we performed using SCANPY[66] (v1.9.8). First, we excluded outlier cells (range [Q1 − 3(Q3 − Q1), Q3 + 3(Q3-Q1)], with Q1 as the lower quartile and Q3 as the upper quartile) in terms of the number of genes, total counts, and percentage of mitochondrial genes. Next, we removed doubly labeled cells using Scrublet[67] (v0.2.3). After filtering 11,780 cells, 88,622 cells were retained. The integration method to remove single-cell platforms and dataset-specific batch effects was performed using Harmony[68] using individuals and batches with normalized gene expression. To annotate major cell types and subtypes based on previously published single-cell RNA sequencing data, annotations of major brain cell types (previously defined by the Allen Brain Institute, https://portal.brain-map.org/atlases-and-data/rnaseq/human-multiple-cortical-areas-smart-seq) were projected onto this study.

Cis-eQTL mapping and COLOC

To test for cis-eQTLs, we used the tensorQTL[69] v.1.0.2 cis_nominal mode with genotypes and a gene expression matrix. Individuals with fewer than 10 cells for each major cell type were filtered out. The pseudo-bulk gene expression matrices were averaged across all counts for each gene in each cell type. As input covariates for the analysis, we included PEER[70] factors 30–70 for each brain cell type and the first four PCA of the genotypes. Each SNP–gene pair used a 1 Mb window within the transcription start site of a gene. We performed TensorQTL cis permutations, with 1,000 permutations per gene. The COLOC package (v3.2-1) was used to assess whether SNPs from the GWAS co-localized with bulk RNA-seq or single-nucleus RNA-seq expression QTLs[71,72]. We extracted a significant genome-wide locus within 1 Mb on either side of the lead SNP (2 Mb wide region total) in the GWAS. In each QTL dataset, we filtered all SNPs of each gene matched with a significant genome-wide locus within 100 kb to test for co-localization. Missing minor allele frequencies were replaced with reference values from the European superpopulation of the 1000 Genomes Project (Phase 3). Matching sets of colocalized SNPs were compared using their P-values. Colocalization was considered when the posterior probability for colocalization (PP.H4) exceeded 0.5, and the eQTL P-value was below 1×10^{-4}. The 95% credible set consisted of the smallest subset of SNPs with a cumulative SNP.PP.H4 of 95%[73].

Analysis of rare noncoding variants

To investigate the association of rare noncoding variants with AD, we employed CWAS, a method originally developed for ultra-rare de novo variants and optimized for cases where each variant appeared in only to 1–3 samples[24,25]. This optimization guided us to set an MAF threshold of 0.1% to effectively capture these ultra-rare variants. We used 19,266,739 rare heterozygous variants (which passed QC criteria as described above) with MAF ≤ 0.001, gnomAD[74] (v3.1) MAF ≤ 0.001 in the non-psychiatric disease subset, TOGO[75] (GEM Japan Whole Genome Aggregation (GEM-J WGA) Panel) MAF ≤ 0.001, and KOVA[76] MAF ≤ 0.001.

Rare noncoding variants were analyzed using CWAS with the CWAS-Plus package v1.2[77] to integrate AD-related functional data and select rare noncoding variants that were strongly associated with AD risk. The framework incorporates diverse epigenomic and transcriptomic data to prioritize rare noncoding variants. The package is available at [https://github.com/joonan-lab/cwas]. CWAS categorizes variants by assessing combinations of five types of annotations: variant type, gene set, functional score, genomic region, and functional annotation. The five annotations were as follows: (1) variant type: variant type based on their length as an SNV or indel; (2) gene set: variants located within the same gene groups; (3) functional score: based on conservation and constraint scores; (4) genomic region: specifying the genomic region where the variant was located; and (5) functional annotation: functional elements to which each variant belonged. Association tests were conducted to determine whether variants belonged to a particular category in each sample.

For the gene set, cell type marker genes from PsychENCODE[78], AD-related pathway genes from a previous study[3], and genes from the tau protein binding Gene Ontology Term (GO:0048156) were used. For the functional score, PhastCons46way[79] (> 2), Phlop46way[80] (> 0.2), and JARVIS[81] (> 0.99) were used. For functional annotation, cell type-specific regulatory elements from single-cell data obtained from postmortem AD brain samples[26,27] (available at https://personal.broadinstitute.org/bjames/AD_snATAC/ and https://compbio.mit.edu/microglia_states/) and AD-specific epigenomic histone acetylation data from postmortem human brains[82] (available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130746) were used (Supplementary Data 3a). Each variant was annotated using VEP[46] (v105). Categories for each variant were created based on a combination of the five annotations. Variants were assigned to multiple categories based on their annotations.

We performed a two-sided binomial exact test within each category to evaluate case–control associations based on the number of samples carrying variants in each category. *P*-values were calculated using a binomial test by comparing the observed case–control proportions within a category with the null expectation (Supplementary Fig. 5e). Relative risk (RR) was computed as the ratio of the proportion of cases to controls carrying the variants. To control for multiple testing in our burden analysis, we estimated the effective number of tests. This was achieved by transforming the correlation structure among categories into a negative Laplacian form. Specifically, we took the absolute values of the correlation matrix and computed a degree matrix by summing correlations for each category. These values were then normalized by the squared entries of the degree matrix to produce a Laplacian matrix. We performed eigen decomposition on this matrix and counted the number of eigenvalues that cumulatively explained at least 99% of the total variance. This number was used to approximate the number of independent tests and guide genome-wide significance thresholds. Next, we assessed whether the nominally significant categories within each genomic region were statistically enriched. Categories with *P*-values below the nominal threshold ($p < 0.05$) were identified and compared across the genomic regions. To confirm the significance of these counts, label-swapped permutation tests were performed, in which phenotypic labels were randomly reassigned across 1,000 iterations to generate a null distribution for the number of significant categories. The observed counts of the significant categories were then compared to this null distribution to identify regions that exhibited enrichment beyond random expectations, confirming the presence of case-enrichment. Only categories with sample counts of 9 or more were included. For power estimation in our CWAS analyses, we performed a logit-based binomial power analysis across a range of hypothetical sample sizes (0 to 70,000) to assess the sensitivity in detecting differences in variant burden between groups. Specifically, for each sample, the variant burden was computed as the total number of positive variant calls and then aggregated to obtain group-level summary statistics (mean, standard deviation, and total counts). A one-tailed binomial test was subsequently applied to determine whether the cumulative variant burden in cases exceeded that in controls at a significance level of 0.05. For the Aβ-positivity test ($n = 925$ for Aβ-positive; $n = 634$ for Aβ-negative), the estimated power was 0.85, while for the diagnosis test ($n = 314$ for DAT; $n = 655$ for CU), the estimated power was 0.098. *n* refers to the number of individual samples.

We employed the DAWN hidden Markov random field model[24,83] to examine genetic factors associated with AD risk within intergenic regions. A network was constructed based on the correlation of sample counts across 2074 intergenic categories selected from a total of

29,917 categories that had 10 or more samples. Clusters of intergenic categories were identified using K-means clustering ($K = 194$), which was determined to be optimal based on the silhouette coefficient, reflecting the correlation structure among intergenic categories. A sparse PCA was performed to estimate the RR for each cluster. Permutation-based z-scores were calculated to standardize the observed metrics across categories. A hidden Markov random field model was then applied to adjust these estimates, considering the enrichment of neighboring clusters within the simulated correlation network. This process yields the Bayesian posterior probabilities for each cluster. Significant clusters were identified using a Bayesian false discovery rate (FDR) threshold of 0.05, applied to posterior probabilities. Clusters meeting this criterion were considered statistically significant. Among the significant clusters, those with an RR greater than 1 were further defined as risk clusters. The CWAS analysis commands and scripts used in this study are available in the GitHub repository (https://github.com/wonlab101/K-ROAD_Alzheimer_WGS). To examine the annotation terms that define five risk clusters, the overlap between each pair of variant categories was quantified to calculate the correlation matrix. Overlap was determined as the count of shared variants between two categories divided by the geometric mean of each category's total number of variants. The correlation values ranged from 0 to 1, where 0 indicated no overlap and 1 indicated complete overlap.

To investigate phenotypic associations, we focused on Aβ-positive samples and divided them into carrier and non-carrier groups for each identified risk cluster. We conducted a linear regression analysis to evaluate the associations between the four risk clusters (Clusters 25, 98, 103, and 194) and four phenotypic variables (verbal memory, visual memory, CDRSB, and K-MMSE), resulting in a total of 16 tests. To assess whether carriers exhibited increased severity in cognitive function phenotypes, we compared the clinical scores of carriers and non-carriers of each risk cluster variant within Aβ-positive samples. Each cluster variable was converted to a factor, and linear regression models were used for each phenotype, adjusting for education, age, and sex as covariates. Although none of the 16 tests met the threshold for significance after FDR correction, our primary goal was to identify variants in the genomic regions with the strongest nominal associations with phenotypes. Based on this criterion, we prioritized the variants in Cluster 25 (C25), which showed the most significant nominal p-values for follow-up analysis. We investigated the interactions between 63 Aβ-positive-only variants in C25, selected based on the criterion of each variant being present in at least two Aβ-positive samples, using Hi-C data via 3DIV[28] (http://3div.kr/hic). We identified protein-coding genes within 2 Mb with a distance-normalized interaction frequency > 2, and confirmed interactions with 109 genes for 39 variants. We examined the excitatory subtype CREs for each variant and determined whether their target genes were significantly downregulated with cognitive impairment in the same excitatory subtype (Exc L2-3 CBLN2 LINC02306), using cell type-specific DEGs described by Mathys et al.[18] from the ROSMAP cohort were used. DecontX- and RUVr-adjusted DEGs were obtained from the AD/Aging Brain Atlas (https://compbio.mit.edu/ad_aging_brain/).

Copy number variant analysis

For CNV (copy number variant) discovery, Parliament2[84] was performed. It runs four callers in parallel: Manta[85], Lumpy[86], Delly[87], and CNVnator[88]. The results from each caller were integrated at the individual level using SURVIVOR[89] and validated using SVTyper[90]. CNVs were filtered based on the following three criteria: (1) deletions and duplications; (2) CNVs detected by Manta; and (3) a maximum CNV length of 1 Mbp. The CNVs identified in each sample were merged into a population set using the SURVIVOR software. The distance between the breakpoints was set to 1000 bp, with a minimum length threshold of 50 bp. The population CNV set was annotated using AnnotSV[91], and only the results from the full-mode annotation were used for the analysis (Supplementary Fig. 19a).

We performed logistic regression to examine the effect of each CNV on Aβ positivity and clinical diagnosis (Supplementary Fig. 5f). CNVs were categorized as carriers or non-carriers. Age, sex, and the three PCs were included as covariates. For the standard regression analysis, CNVs identified in only one patient were excluded. The results of the regression analysis were adjusted by multiple comparisons using the Bonferroni method, FDR, and the Bonferroni method, which is based on cytobands. Significant CNVs were validated by visualization using samplot[92].

Short tandem repeat analysis

We identified short tandem repeats (STR) using ExpansionHunter[93] (v5.0.0) with default parameters on a panel of 312,302 polymorphic STRs from Guo et al.[13]. The three analytical models for STR analysis were extended and adapted from the framework described by Guo et al[13]. For quality control, one sample with poor calling quality was excluded. We also performed PCA using the R package stats (v4.2.2) based on STR coverage and excluded samples that differed in PC1 and PC2 distributions by more than four times the SD from the mean. In addition, samples outside the main cluster were filtered using a threshold of less than 220 on PC2 (Supplementary Fig. 20). We excluded 27,551 STRs located in segmentally duplicated regions, leaving 293,751 STRs for further analysis. We calculated the MAF for each STR to examine its distribution (Supplementary Fig. 19b). All alleles were included in the STR analysis. The remaining 293,751 STRs were subjected to multiple hypothesis testing corrections using the Bonferroni method, with a significance threshold of 1.70×10^{-7} (0.05/293,751). In all logistic regression analyses, the following covariates were included: sex, age, sequencing batch, PC 1–3, average STR coverage per sample, and STR coverage. We excluded the coverage for each STR when comparing the samples.

We assessed the association between STR genotype and AD risk using a dominant model and analyzed the length of the two alleles for each STR. Because of the non-normal distribution of many STR genotypes, we conducted a rank-based inverse normal transformation before analysis. We performed logistic regression to evaluate the association between STR and disease status. The inverse-normal transformed allele count of each STR locus was used as the primary predictor. Models were adjusted for sex, age, sequencing batch, sequencing depth, *APOE* ε4 carrier status, and PC1-3 (Supplementary Fig. 5g, h). Analyses were conducted both with and without adjustment for APOE ε4 carrier status. We compared the P-values from logistic regression on transformed STR genotypes with those from non-transformed models to determine absolute effect sizes. We confirmed the similarity of P-values before and after transformation (correlation Pearson $r^2 = 0.90$, $p < 2.2 \times 10^{-16}$).

We conducted a burden test to compare the number of STR expansions between the cases and controls. For the panel of 293,751 STRs, expansions were defined as STR lengths ≥ 5, 10, or 20 repeat units longer than the GRCh38 reference. We constructed 2×2 contingency tables for each STR and used Fisher's exact test to evaluate the differences in expansion burden. Bonferroni correction was used for multiple hypothesis testing, with a significance threshold of 1.70×10^{-7}. In addition, we analyzed expansions observed in one, ≤ 5, ≤ 10, or ≤ 100, or without frequency cutoff, comparing total expansions between cases and controls.

We used the density-based spatial clustering of applications with noise (DBSCAN) outlier detection method (adapted from previous STR studies[13,94]) to identify extreme STR tract lengths in 1515 samples. DBSCAN, an unsupervised clustering technique, determines outlier clusters by examining data density[95]. We established clusters by setting criteria for the minimum number of reachable data points (μ) within a maximum distance (ε). Outliers representing extreme STR lengths

were identified as data points that could not be reached by the clusters. Specifically, we set ε as twice the mode of STR lengths, and μ as $\log_2$ of the sample size. Each STR was analyzed by inputting the length of the two alleles per individual into DBSCAN. Before the analysis, we conducted a linear regression analysis to account for sample and technical covariates. The residuals from this regression were used as inputs for DBSCAN. Next, we counted the number of outlier STRs per sample and measured the odds ratios based on each threshold for Aβ positivity. We examined the differences in Aβ levels among groups using a linear model, with covariates consistent with those used in the logistic regression model.

Three SNPs, rs28372356, rs6923619, and rs78009495, which were identified in our GWAS as having significant and suggestive associations, were evaluated for LD with STR outliers. A total of 15,910 STR outliers were analyzed, focusing specifically on 84 unique STRs located within a 2 Mb window of these SNPs to assess potential LD relationships. The LD calculations were performed using PLINK version 1.9. The merged VCF file containing both the lead SNPs and STRs was first converted to a PLINK binary format (BED, BIM, and FAM files). For each of the three lead SNPs, pairwise LD (r^2) values were computed with STR outliers within the defined 2 Mb window to identify SNP-STR associations in close proximity. An r^2 value greater than 0.1 was considered indicative of meaningful LD between the SNP and the corresponding STR.

Gene set enrichment analysis was performed using the ClusterProfiler package (v4.6.2) in R software. Our analysis compared STRs found in samples with 40 or more outliers ($n = 4424$) with the entire reference STR panel ($n = 293,751$). Each STR was linked to a gene with the nearest transcription start site within 500 kb. To assess enrichment, we employed the enrichGO function in ClusterProfiler with the following parameters: keyType = ENTREZID, ont = ALL, pvalueCutoff = 0.05, and qvalueCutoff = 0.05.

PRS calculation

PRS was calculated using European-based GWAS data and PRS-cs[96,97] (v1.0.0). Two GWAS studies were employed for PRS calculation, one based on clinical diagnosis[3] ($n = 487,511$) and the other on Aβ positivity[15] ($n = 11,816$). The region near *APOE* was excluded from the analysis (chromosome 19, 43895848 to 45996742/GRCh38[98]). For the PRS-cs analysis, we utilized the UK Biobank European LD reference and set the global shrinkage parameter (phi) to 1e-02, which is appropriate for highly polygenic traits. PRS scoring was conducted using the score option in PLINK software (v1.90), incorporating a total of 663,786 and 579,170 variants derived from the clinical diagnosis and Aβ GWAS, respectively. In addition, we calculated the PRS using three lead SNPs (rs28372356, rs6923619, and rs78009495) identified in this study. The score function in PLINK v1.90[44] was used to compute the PRS for each variant using the beta values. Statistical tests were conducted using the R software (v4.3.1).

A regression analysis was performed to assess the association between PRS and AD pathology and cognitive function. AD pathology was evaluated using clinical diagnosis (CU versus DAT) and Aβ positivity, with logistic regression and adjustments for age, sex, 10 PCs of genetic ancestry, batch, and education. Cognitive function was evaluated using the K-MMSE, CDR-SB, visual memory, and verbal memory tests, with linear regression adjustments for age, sex, 10 PCs of genetic ancestry, batch, and education.

Comparing genetic factors

The explanatory power of the identified loci was calculated using incremental r^2 values[99,100]. Nagelkerke's r^2 was computed using logistic regression, and the differences in r^2 between the null model and the models with variants were calculated as incremental r^2. The null model included covariates and *APOE* genotypes (age, sex, batch effects, *APOE* ε4 genotype, *APOE* ε2 genotype, and PC 1 to 10). First, we integrated lead variants identified in the European GWAS[3] (Bellenguez et al.) into the model. A total of 67 variants were identified in the Korean genotype data. Second, we added the genotypes of one significant lead variant and two suggestive lead variants identified in this study. Finally, the RVs and SVs identified in this study were included. These models are defined as follows:

Null model: clinical diagnosis or Aβ ~ *APOE* ε4 genotype + *APOE* ε2 genotype + covariates (age, sex, batch effects, PC 1 to 10).

Model 1 (lead variants identified in Europeans): clinical diagnosis or Aβ ~ 67 lead variants identified in the European GWAS + null model.

Model 2 (lead variants identified in Europeans and Koreans): clinical diagnosis or Aβ ~ 3 lead variants identified in the Korean GWAS + model 1.

Model 3 (lead variants identified in Europeans and Koreans, along with rare and SV from the Korean cohort): clinical diagnosis or Aβ ~ rare coding variants (DRC7) + rare noncoding variants (C25) + STR + CNV + model 2.

To confirm the phenotypic associations of genetic factors, we examined the genetic effects of *APOE* ε4 carriers, PRS top 20%, rare coding variants carriers, and SV carriers. The top 20% of PRS were calculated based on Aβ GWAS data ($n = 11,816$). Rare noncoding variant carriers were defined as individuals who possessed at least one rare noncoding variant from C25. We excluded carriers of rare coding variants because these variants exhibited a protective effect contrary to the effects observed for other variants. SV carriers were defined as individuals with either CNVs in the *HPSE2* gene region or an STR expansion of 40 or more repeats.

Samples with missing data for any of the categories (*APOE* ε4 carrier, PRS top 20%, rare noncoding variants, and SVs) were excluded from the analysis, resulting in 1,495 individuals analyzed. We categorized individuals into eight groups, including the non-carrier group, groups with individuals carrying only one of the genetic factors (*APOE* ε4 only, PRS only, rare noncoding variants only, and SV only), as well as groups with two, three, and four factors. The group carrying all four factors was excluded from further analysis because of the presence of only one sample. First, we compared the non-carrier group with other groups and then compared the *APOE*-only group with the two- or three-factor groups. The Wilcoxon rank-sum test was used to test the statistical significance. The Bonferroni correction was applied to adjust for multiple comparisons.

Next, we performed analyses on groups containing individuals who were carriers of both *APOE* ε4 and high PRS, rare noncoding variants, or SV, and compared them to the group comprising only carriers of *APOE* ε4. For this analysis, we used the PRS calculated with the three lead SNPs identified in the K-ROAD study. Group-wise significance comparisons were conducted using linear regression, adjusting for age, sex, PC 1–10, batch, and education. Statistical tests were performed using the R software (version 4.3.1).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The GWAS summary statistics for the Korean cohort are available in the NHGRI-EBI GWAS Catalog under accession numbers GCST90566388 and GCST90566389 [https://www.ebi.ac.uk/gwas]. The Japanese GWAS summary statistics used in this study were provided by Prof. Daichi Shigemizu. The single-nuclei RNA-seq data for the Korean participants are accessible for collaborative research under restricted conditions to protect participant privacy. For inquiries regarding data access, please contact the corresponding author (S.W. Seo (sangwonseo@empas.com)). Data access requests will be reviewed and responded to within 2 to 4 weeks of receipt. The summary statistics of cell type-specific eQTLs (https://zenodo.org/

records/7276971), MiGA eQTLs (https://doi.org/10.5281/zenodo.4118605), multi-ethnic eQTLs (http://icahn.mssm.edu/brema), and MetaBrain eQTLs (https://www.metabrain.nl/) are available in public repositories. The results of single-cell analyses from ROSMAP-MIT, including DEGs, are available at the AD/Aging Brain Atlas (https://compbio.mit.edu/ad_aging_brain/). The SEA-AD analysis results are available at the Allen Brain Map (https://portal.brain-map.org/explore/seattle-alzheimers-disease). The cell type-specific regulatory elements derived from single-cell data obtained from postmortem AD brain samples are available at [https://personal.broadinstitute.org/bjames/AD_snATAC/] and [https://compbio.mit.edu/microglia_states/]. AD-specific epigenomic histone acetylation data from postmortem human brains, available at [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130746]. The chromatin interaction data are available at 3DIV. The ExpansionHunter input variant catalogs are available from Guo et al. [https://github.com/mhguo1/AD_STR]. Source data are provided as a Source Data file. Source data are provided in this paper.

Code availability

Publicly available software was used for all analyses. This code is available on GitHub (https://github.com/wonlab101/K-ROAD_Alzheimer_WGS) at https://doi.org/10.5281/zenodo.15189321. CWAS-Plus is available on GitHub (https://github.com/joonan-lab/cwas/tree/CWAS-v.1.2). COLOC is available on GitHub (https://github.com/RajLabMSSM/downstream-QTL/tree/master). The eQTL mapping pipeline is available on GitHub (https://github.com/RajLabMSSM/QTL-mapping-pipeline).

References

1. Collaborators, G. B. D. D. F. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. *Lancet Public Health* **7**, e105–e125 (2022).
2. Gatz, M. et al. Role of genes and environments for explaining Alzheimer disease. *Arch. Gen. Psychiatry* **63**, 168–174 (2006).
3. Bellenguez, C. et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).
4. Wightman, D. P. et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).
5. Baker, E. et al. What does heritability of Alzheimer's disease represent? *PLoS ONE* **18**, e0281440 (2023).
6. Biddanda, A., Rice, D. P. & Novembre, J. A variant-centric perspective on geographic patterns of human allele frequency variation. *Elife* **9**, e60107 (2020).
7. Lake, J. et al. Multi-ancestry meta-analysis and fine-mapping in Alzheimer's disease. *Mol. Psychiatry* **28**, 3121–3132 (2023).
8. Miyashita, A., Kikuchi, M., Hara, N. & Ikeuchi, T. Genetics of Alzheimer's disease: an East Asian perspective. *J. Hum. Genet.* **68**, 115–124 (2023).
9. Asanomi, Y. et al. A rare functional variant of SHARPIN attenuates the inflammatory response and associates with increased risk of late-onset Alzheimer's disease. *Mol. Med.* **25**, 20 (2019).
10. Asanomi, Y. et al. A functional variant of SHARPIN confers increased risk of late-onset Alzheimer's disease. *J. Hum. Genet.* **67**, 203–208 (2022).
11. Park, J. Y. et al. A missense variant in SHARPIN mediates Alzheimer's disease-specific brain damages. *Transl. Psychiatry* **11**, 590 (2021).
12. Cooper, Y. A. et al. Functional regulatory variants implicate distinct transcriptional networks in dementia. *Science* **377**, eabi8654 (2022).
13. Guo, M. H., Lee, W. P., Vardarajan, B., Schellenberg, G. D. & Phillips-Cremins, J. E. Polygenic burden of short tandem repeat expansions promotes risk for Alzheimer's disease. *Nat. Commun.* **16**, 1126 (2025).
14. Dubois, B., von Arnim, C. A. F., Burnie, N., Bozeat, S. & Cummings, J. Biomarkers in Alzheimer's disease: role in early and differential diagnosis and recognition of atypical variants. *Alzheimers Res. Ther.* **15**, 175 (2023).
15. Ali, M. et al. Large multi-ethnic genetic analyses of amyloid imaging identify new genes for Alzheimer disease. *Acta Neuropathol. Commun.* **11**, 68 (2023).
16. Shigemizu, D. et al. Whole-genome sequencing reveals novel ethnicity-specific rare variants associated with Alzheimer's disease. *Mol. Psychiatry* **27**, 2554–2562 (2022).
17. Shigemizu, D. et al. Ethnic and trans-ethnic genome-wide association studies identify new loci influencing Japanese Alzheimer's disease risk. *Transl. Psychiatry* **11**, 151 (2021).
18. Mathys, H. et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell* **186**, 4365–4385 (2023).
19. Beker, M. & Kılıç, E. The role of circadian rhythm in the regulation of cellular protein profiles in the brain. *Turk. J. Med. Sci.* **51**, 2705–2715 (2021).
20. Ding, W. et al. Adaptive functions of structural variants in human brain development. *Sci. Adv.* **10**, eadl4600 (2024).
21. Gastfriend, B. D. et al. Wnt signaling mediates acquisition of blood-brain barrier properties in naïve endothelium derived from human pluripotent stem cells. *Elife* **10**, e70992 (2021).
22. Kirmiz, M., Vierra, N. C., Palacio, S. & Trimmer, J. S. Identification of VAPA and VAPB as Kv2 Channel-Interacting Proteins Defining Endoplasmic Reticulum-Plasma Membrane Junctions in Mammalian Brain Neurons. *J. Neurosci.* **38**, 7562–7584 (2018).
23. Kim, S. et al. An integrative single-cell atlas for exploring the cellular and temporal specificity of genes related to neurological disorders during human brain development. *Exp. Mol. Med.* **56**, 2271–2282 (2024).
24. An, J. Y. et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018).
25. Werling, D. M. et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50**, 727–736 (2018).
26. Xiong, X. et al. Epigenomic dissection of Alzheimer's disease pinpoints causal variants and reveals epigenome erosion. *Cell* **186**, 4422–4437 (2023).
27. Sun, N. et al. Human microglial state dynamics in Alzheimer's disease progression. *Cell* **186**, 4386–4403.e29 (2023).
28. Kim, K. et al. 3DIV update for 2021: a comprehensive resource of 3D genome and 3D cancer genome. *Nucleic Acids Res.* **49**, D38–D46 (2021).
29. Jung, S.-H. et al. Transferability of Alzheimer disease polygenic risk score across populations and its association with Alzheimer disease-related phenotypes. *JAMA Netw. Open* **5**, e2247162–e2247162 (2022).
30. Eschbach, J. & Dupuis, L. Cytoplasmic dynein in neurodegeneration. *Pharm. Ther.* **130**, 348–363 (2011).
31. Kimura, N., Okabayashi, S. & Ono, F. Dynein dysfunction disrupts β-amyloid clearance in astrocytes through endocytic disturbances. *Neuroreport* **25**, 514–520 (2014).
32. Lee, W. P. et al. Association of common and rare variants with Alzheimer's disease in more than 13,000 diverse individuals with whole-genome sequencing from the Alzheimer's Disease sequencing project. *Alzheimers Dement* **20**, 8470–8483 (2024).
33. Zhao, S. et al. A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. *Nat. Genet.* **55**, 346–354 (2023).

34. Christensen, K. J., Multhaup, K. S., Nordstrom, S., Voss, K.J.P.A.A.J.o.C. & Psychology, C. A cognitive battery for dementia: Development and measurement characteristics. *Psychol. Assess. J. Consult. Clin. Psychol.* **3**, 168 (1991).
35. Ahn, H. J. et al. Seoul Neuropsychological Screening Battery-dementia version (SNSB-D): a useful tool for assessing and monitoring cognitive impairments in dementia patients. *J. Korean Med. Sci.* **25**, 1071–1076 (2010).
36. Albert, M. S. et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement* **7**, 270–279 (2011).
37. McKhann, G. M. et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 263–269 (2011).
38. Klunk, W. E. et al. The Centiloid Project: standardizing quantitative amyloid plaque estimation by PET. *Alzheimers Dement* **11**, 1–15 e1-4 (2015).
39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
40. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
41. McKenna, A. et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
42. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
43. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11 10 1–11 10 33 (2013).
44. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
45. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
46. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
47. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
48. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* **32**, 894–899 (2011).
49. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* **12**, 103 (2020).
50. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
51. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
52. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
53. John Lee, J. L., Xiuwen Zheng ggmanh: Visualization Tool for GWAS Result. *R package version 1.8.0* (2024).
54. Wilke, C. O. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. (2024).
55. Myers, T. A., Chanock, S. J. & Machiela, M. J. LDlinkR: An R package for rapidly calculating linkage disequilibrium statistics in diverse populations. *Front. Genet.* **11**, 157 (2020).
56. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
57. Rosenbloom, K. R. et al. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* **41**, D56–D63 (2013).
58. Bryois, J. et al. Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat. Neurosci.* **25**, 1104–1112 (2022).
59. Lopes, K. P. et al. Genetic analysis of the human microglial transcriptome across brain regions, aging and disease pathologies. *Nat. Genet.* **54**, 4–17 (2022).
60. Zeng, B. et al. Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. *Nat. Genet.* **54**, 161–169 (2022).
61. de Klein, N. et al. Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nat. Genet.* **55**, 377–388 (2023).
62. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
63. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*, (Springer-Verlag New York, 2016).
64. Gabitto, M. I. et al. Integrated multimodal cell atlas of Alzheimer's disease. *Res Sq* (2023).
65. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
66. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
67. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291 (2019).
68. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
69. Taylor-Weiner, A. et al. Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).
70. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
71. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
72. Fujita, M. et al. Cell subtype-specific effects of genetic variation in the Alzheimer's disease brain. *Nat. Genet.* **56**, 605–614 (2024).
73. Cho, C. et al. Large-scale cross-ancestry genome-wide meta-analysis of serum urate. *Nat. Commun.* **15**, 3441 (2024).
74. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
75. Mitsuhashi, N. et al. TogoVar: A comprehensive Japanese genetic variation database. *Hum. Genome Var.* **9**, 44 (2022).
76. Lee, J. et al. A database of 5305 healthy Korean individuals reveals genetic and clinical implications for an East Asian population. *Exp. Mol. Med.* **54**, 1862–1871 (2022).
77. Kim, Y. et al. CWAS-Plus: estimating category-wide association of rare noncoding variation from whole-genome sequencing data with cell-type-specific functional data. *Brief. Bioinform.* **25**, bbae323 (2024).
78. Psych, E. C. et al. The PsychENCODE project. *Nat. Neurosci.* **18**, 1707–1712 (2015).
79. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
80. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Re.s* **20**, 110–121 (2010).
81. Vitsios, D., Dhindsa, R. S., Middleton, L., Gussow, A. B. & Petrovski, S. Prioritizing non-coding regions based on human genomic

constraint and sequence context with deep learning. *Nat. Commun.* **12**, 1504 (2021).
82. Nativio, R. et al. An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease. *Nat. Genet.* **52**, 1024–1035 (2020).
83. Liu, L. et al. DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. *Mol. Autism* **5**, 22 (2014).
84. Zarate, S. et al. Parliament2: Accurate structural variant calling at scale. *Gigascience* **9**, giaa145 (2020).
85. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
86. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
87. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
88. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
89. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
90. Chiang, C. et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
91. Geoffroy, V. et al. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
92. Hall, J. E., Hofman, W. F. & Ehrhart, I. C. Venous occlusion pressure and vascular permeability in the dog lung after air embolization. *J. Appl Physiol.* **65**, 34–40 (1988).
93. Dolzhenko, E. et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).
94. Trost, B. et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* **586**, 80–86 (2020).
95. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*. **96**, 226–231 (1996).
96. Ge, T., Chen, C. Y., Ni, Y., Feng, Y. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
97. Kachuri, L. et al. Principles and methods for transferring polygenic risk scores across global populations. *Nat. Rev. Genet.* **25**, 8–25 (2024).
98. Escott-Price, V. et al. Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain* **138**, 3673–3684 (2015).
99. Kim, B. et al. Mapping and annotating genomic loci to prioritize genes and implicate distinct polygenic adaptations for skin color. *Nat. Commun.* **15**, 4874 (2024).
100. Turley, P. et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).

Acknowledgements

We thank the National Center for Geriatrics and Gerontology (NCGG) investigators for providing the GWAS summary data analyzed from the WGS data. GWAS summary statistics were provided by Prof. Daichi Shigemizu. In addition, we thank Prof. Carlos Cruchaga for providing the Aβ GWAS summary statistics of a non-Hispanic white cohort. This study utilized BeauBrain Amylo's image processing technology to quantify amyloid uptakes using PET-CT. This study was supported by Future Medicine 2030 Project of the Samsung Medical Center (SMX1250081); the Korea Dementia Research Project through the Korea Dementia Research Center (KDRC), funded by the Ministry of Health & Welfare and Ministry of Science and ICT, Republic of Korea (RS-2020-KH106434, RS-2022-KH125557); the National Research Foundation of Korea (NRF) (RS-2019-NR040057, RS-2023-00262527, RS-2023-00247408, RS-2025-00553304); Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2021-II212068, Artificial Intelligence Innovation Hub); Korea University (K2501651); and Korea National Institute of Health (2024-ER1003-01, 2023-ER1001-02).

Author contributions

Conceptualization and design: J.P. Kim, M. Cho, C. Kim, H. Lee, S.W. Seo, J.Y. An, and H.-H. Won; Data acquisition, analysis, or interpretation of data: J.P. Kim, M. Cho, C. Kim, H. Lee, B. Jang, S.-H. Jung, Y. Kim (Korea University), I.G. Kim, S. Kim, D. Shin, E.H. Lee, J.-Y. Lee, Y.C. Park, H. Jang, B.-H. Kim, H. Ham, B. Kim, and Y. Kim (Sungkyunkwan University), A.H. Cho, T. Raj, H.J. Kim, D.L. Na, S.W. Seo, J.Y. An, and H.-H. Won; Statistical analysis: J.P. Kim, M. Cho, C. Kim, H. Lee, B. Jang, J.Y. An, and H.-H. Won; Drafting of the manuscript: J.P. Kim, M. Cho, C. Kim, H. Lee, B. Jang, S.W. Seo, J.Y. An, and H.-H. Won; Critical revision of the manuscript: J.P. Kim, M. Cho, C. Kim, H. Lee, B. Jang, S.-H. Jung, Y. Kim (Korea University), I.G. Kim, S. Kim, D. Shin, E.H. Lee, J.-Y. Lee, Y.C. Park, H.Jang, B.-H.Kim, H. Ham, B. Kim, Y. Kim (Sungkyunkwan University), A.H. Cho, T. Raj, H.J. Kim, D.L. Na, S.W. Seo, J.Y. An, and H.-H. Won; Supervision: S.W. Seo, J.Y. An, and H.-H. Won; Funding acquisition: S.W. Seo, J.Y. An, and H.-H. Won.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-59949-y.

Correspondence and requests for materials should be addressed to Sang Won Seo, Joon-Yong An or Hong-Hee Won.

Peer review information *Nature Communications* thanks Michael Guo, Xushen Xiong, who co-reviewed with Jiuhong Nan, and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

[1]Alzheimer's Disease Convergence Research Center, Samsung Medical Center, Seoul, Republic of Korea. [2]Department of Neurology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea. [3]Neuroscience Center, Samsung Medical Center, Seoul, Republic of Korea. [4]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology (SAIHST), Sungkyunkwan University, Seoul, Republic of Korea. [5]Department of Integrated Biomedical and Life Science, Korea University, Seoul, Republic of Korea. [6]L-HOPE Program for Community-Based Total Learning Health Systems, Korea University, Seoul, Republic of Korea. [7]Department of Health Sciences and Technology, Samsung Advanced Institute for Health Sciences & Technology (SAIHST), Sungkyunkwan University, Seoul, Republic of Korea. [8]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [9]Nash Family Department of Neuroscience & Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [10]Ronald M. Loeb Center for Alzheimer's Disease, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [11]Department of Medical Informatics, Kangwon National University College of Medicine, Chuncheon, Republic of Korea. [12]Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN, USA. [13]Indiana Alzheimer Disease Research Center, Indiana University School of Medicine, Indianapolis, IN, USA. [14]Oneomics Co., Ltd, Gyeonggi-do, Republic of Korea. [15]Division of Bio Bigdata, Department of Precision Medicine, Korea National Institution of Health, Cheongju, Republic of Korea. [16]Department of Neurology, Seoul National University Hospital, Seoul National University School of Medicine, Seoul, Republic of Korea. [17]Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [18]Estelle and Daniel Maggin Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [19]School of Biosystem and Biomedical Science, College of Health Science, Korea University, Seoul, Republic of Korea. [20]Samsung Genome Institute, Samsung Medical Center, Seoul, Republic of Korea. [21]These authors contributed equally: Jun Pyo Kim, Minyoung Cho, Chanhee Kim. [22]These authors jointly supervised this work: Sang Won Seo, Joon-Yong An, and Hong-Hee Won. ✉e-mail: sangwonseo@empas.com; joonan30@korea.ac.kr; wonhh@skku.edu