nature portfolio

Peer Review File

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

Corresponding Author: Professor Hong-Hee Won

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In this paper by Kim et al, the authors perform a whole genome sequencing analysis of a cohort from Korea with and without Alzheimer's disease. The authors perform comprehensive analyses across multiple forms of genetic variation. The authors also explore the association between genetic variation and AD-related endophenotypes, including amyloid positivity and cognitive testing scores. The authors identify several new statistical associations, though many of these are only nominally significant. The authors also recapitulate the results of a recent study examining short tandem repeats in AD. The analyses are rigorous and comprehensive and overall the manuscript is clear. However, there are some important concerns related to the manuscript.
- In Extended Figures 3A & D, there appear to be a few individuals who were classified as DAT, but have negative AB biomarkers. Can the authors explain this?
- For the APOE associations, there is some emerging evidence that there are other alleles near APOE outside of the traditional APOE2/3/4 that may also be associated with AD. It would be interesting if the authors could perform conditional analyses to see if there are additional independent SNP association signals outside of the E2/3/4 alleles.
- For the APCDD1 SNP association, can the authors discuss why this was not uncovered in other AD studies despite the modest size of this study? Is it because the alleles are much more common in the Korean population?
- For the single variant associations at APCDD1, SAMD3, and PTPRD, the authors should perform statistical fine-mapping to see if they can refine the credible set of potentially causal variants that may aide in the identification of the candidate gene at each of these loci.
- For DCR7 rare variant association, has there been any suggestive signal for this gene in other AD genetics papers? Are the alleles seen in DCR7 in this work present in other populations?
- In Figures 2b-c, why did the authors not show the eQTL locus zoom plot using ROSMAP or GTEx data as was used for statistical colocalization (as in Figure 2D)? Similarly, why did they not perform colocalization analyses for the Metabrain QTLs?
- In the CWAS study, the authors say that the use Lasso regression to prioritize noncoding categories contributing to AB positivity. The authors should better explain how this was done, as I am quite confused by this section.
- For figure 4j, the authors should provide an additional zoomed in version (e.g., for outlier counts < 100), as it is hard to see the distribution.
- For the risk score analysis, heterozygous APOE4 should not be considered as "monogenic". I would generally avoid the term "monogenic" here since it implies that all APOE4 carriers have AD, which would obviate the entire analysis.
- Throughout, the authors should generate quantile-quantile plots (or suitable alternative) to test for inflation of test statistics. This should be done for at least the GWAS, rare variant analysis, CWAS, and STR analysis under Model 1.
- Throughout, the authors need to more clearly delineate how many hypotheses are being test. For the analyses related to Figure 3g and 3h, these do not appear to corrected for multiple hypothesis testing.

Minor:
- In line 161, it should be TMEM200A, not TEME200A.
- In line 201, the authors suggest that they are performing "functional" analyses. I do not think they are performing functional analyses, but rather statistical colocalizations. I would recommend removing the term "functional analyses"

Reviewer #2

(Remarks to the Author)
In this manuscript, Won and colleagues carried out a GWAS analysis based on the WGS datasets across ~1600 Korean individuals with and without AD. They identified common and rare variants, as well as structural variants (SVs) that are potentially associated with clinical diagnosis of AD and a-beta aggregation. Overall, the genes and loci identified by this study could potentially enhance our understanding of the genetic regulation of AD and help explain the missing variance of AD. However, there are multiple issues with the current manuscript such as its novelty over the existing Asian-based AD GWASs, and the reliability of the genetic association identified, especially regarding the rare variants and SVs. Moreover, the results are generally hard to follow because of the lack of technical/methodological details. The specific comments are as follows:

Major comments:
1. Since there have been multiple studies focusing on the AD GWAS analysis based on Asian populations (https://www.nature.com/articles/s10038-022-01050-z), which showed a comparable or ever larger sample size compared to this manuscript, what are the advantages of the current study? Also, it's important to carry out a comparison with the existing Asian AD GWASs and clarify which loci are reproducible and which are novel.

2. In the rare variant analysis, there are very few individuals (sometimes only 1) carrying the alternative alleles, and therefore I am worried about the reliability of the results. For example, in gene-based association analysis for DRC7, there are many variants where only one individual carries the alternative allele, and therefore might not be convincing to identify their genetic correlation with AD (line 147, Table S2). If the underlying consideration is to build up the association between the gene, rather than individual genetic variants, then the directionality (risk vs protective) of each variant should be considered.

3. The authors defined common variants using a threshold of MAF>1% while rare variants using MAF<0.1%. Did they exclude the variants with MAF between 0.1% and 1% for analysis? Also, with ~1500 samples, MAF 0.1% means only 1-3 samples with the mutation detected, is it biologically meaningful or statistically feasible to analyze these variants?

4. In the introduction, the authors mentioned that "the current GWASs only account for ~15% of the phenotypic variance", could the authors estimate the additional fraction that this study identified?

5. For the rare variants located in the coding regions, the predicted biological effect of these variants on the function/structure of the corresponding protein should be evaluated, such as using PhyloP conservation score, structural domain analysis, and AlphaFold prediction.

6. Figure 2f,g: Why did the authors observe different directionalities of changes for different AD-related variables? Also, can the authors further illustrate the positive vs negative association effects between these two genes and AD severity by analyzing the effect directionality in eQTL vs in AD GWAS?

7. There is a lack of connection between different sections of the results. The author could add more description about the aims of analysis and the logical connection between results in the main text. For examples:
Please clarify the analysis of Fig. 2d in the main text (lines 201-207). If the authors only focused on the SAMD3 locus, why were the FOSB and APOE presented in the results?
The author should explain why ARHGAP18, PTPRD, EPB41L2, VAPA, and TXNDC2 were included as prioritized genes for functional analysis (lines 227-234).

8. Some Figure/Table legends are incomplete and need to be replenished to increase the readability of the results. There are some examples (but not all of them):
In Fig. 1e, what do the different yellow or grey bars mean?
In Fig. S1 e-h, please add the legends of bar width.
In Fig. 1f, please add the information on how to build a connection between AD-associated variants and candidate genes into the legend. The author should clarify where the DEGs are from, is the data collected by this study (need to add sample size of each group) or previous studies (need to add reference)? I suggest the author provide a table for details of eQTLs (reference, tissue/cell type, p-value and beta) and DEGs (Fold change, p-value, phenotype, reference/this study).
Please add the legends of "variant_type" in Table S2C.

9. In Fig. 1e-f, peak-to-gene linking was built at cell-type resolution, I was wondering whether this connection of AD-associated loci is consistent with cell-type resolved eQTL. Is there colocalization between AD-associated loci and cell-type resolved eQTLs, and whether these colocalized cell types are consistent with the evidence for the scATAC-seq and DEG analysis?

10. Did the author consider the effect directionality (risk vs protective) of AD-associated loci for the genetic analysis in Fig. 5? In Fig. 5b, is there a significant difference between the 'APOE only' group and '2/3 factors' groups? In Fig. 5c,d, the author should provide the results of 'APOE only' group. In Fig. 5b-d, why was the PRS calculated using the European GWAS dataset? I would suggest including the PRS group calculated by all the associated loci detected by this study to the analysis in Fig. 5.

11. The author should provide minor allele frequency distribution of the structural variants identified, and whether rare SVs were excluded or included in the analysis.

Minor comments:

1. In lines 112-114, is the age of the CU group significantly lower than that of the DAT and MCI groups?
2. In Table 1, could the authors add information regarding the associated phenotypes and the distance between lead variants and TSS of candidate genes?

3. Please unify the description of brain cell types throughout the main text.

4. In Fig.1, “B” and “C” should be lowercase (Lines 179-180).

5. In Fig. 3c, intergenic loci were marked in red, but the splice site category was not highlighted (line 284).

6. The authors may consider performing an enrichment analysis of STRs in different genomic regions in addition to the pie chart (Fig. 4d, lines 354-355).

7. Is there a significant correlation between the count of STR outliers and the odd ratio of Aβ-positive per person (Fig. 4k, lines 380-381)?

8. “M” of line 207 should be lowercase.

9. The length of deletion in the HPSE2 region is inconsistent between Fig. 4b (309kb) and Extended Fig.11d.

Reviewer #3

(Remarks to the Author)
Kim et al. report a whole-genome sequencing analysis of Alzheimer's disease (AD) in an East Asian population, analyzing n=1559 individuals from the Korean AD cohort. The potential strength of the study is the comprehensive nature of the genetic scan, analyzing single nucleotide variations (SNVs), gene-based rare variant aggregation tests, and copy-number variation analysis. The short tandem repeat outlier analysis (Fig 4k-n) provides new insights into the field. However, the reviewer has concerns about the statistical rigor of the methodology and the interpretation of the results. The manuscript is overall well written. Overall, I cannot recommend publishing the current version of the manuscript.

[Major points (#1-#9)]

1. The reviewer is concerned about the statistical rigor of the analysis, especially the liberating use of a seemingly arbitrary statistical significance threshold. Table 1, which summarizes the "statistically significant" associations reported in the study, lists only one genetic variant (rs28372356). The other reported associations are "suggestive." The authors select some loci with "suggestive" statistical significance and dive into follow-up analysis, for example, in Figure 2. However, using a "suggestive" threshold for prioritizing loci is not a standard practice. Moreover, the authors needed to properly apply multiple hypothesis testing corrections, adjusting for the number of genetic variants and traits considered in their genetic discovery. As such, analyses on "suggestive" loci/associations should be presented in the supplement, not the main text. Authors should provide compelling evidence that their discovery is more than statistical noise or avoid overinterpreting the subsignificant associations.

2. Did the authors check if genotyping errors or other technical factors do not drive the genome-wide significant associations (rs28372356)? It would be helpful to report genotype call rate, quality scores, and Hardy–Weinberg Equilibrium Test p-values in the supplement.

3. Did the authors perform power calculations? The fine-mapping analysis for amyloid beta in Fig 2 may not have sufficient power, especially when the lead association does not reach the genome-wide significance threshold.

4. Similarly, the category-wide association (CWAS) study in Fig 3 does not meet statistical significance (page 16, line 282) for 6766 repeated tests. Moreover, the 6766 categories are selected using the same set of individuals ("double dipping"). The authors should drop overinterpretation (for example, "tend to be enriched" in line 285). The authors should justify the statistical rigor of the CWAS analysis or drop the analysis from the manuscript.

5. Figure 4. The cytoband-based Bonferonni correction is not explained in the Methods. The false positive discovery for the ZNF7 locus (line 347) is alarming and casts doubts whether the type I error is properly controlled.

6. Fig 4h. For the Model 1 test for STRs in the APOE locus, the authors should present the association results adjusted for the APOE4 carrier status as the primary results. Does "two additional STRs" meet the multiple hypothesis adjusted threshold?

7. Fig 4k-m. The short tandem repeat outlier analysis looks nice. Can authors test whether the identified outliers are in LD with SNVs and if the conditional analysis would attenuate the enrichment observed in Fig 4k-m?

8. Polygenic score analysis from European GWAS data. Can authors comment on whether the genetic discovery from European cohorts is readily applicable to the analysis of East Asian samples? Some results (for example, Supplementary Figure 13d) may not survive the statistical significance after adjusting for the number of tests (number of PRS models and tested traits).

9. Based on the results in Fig 5c-d, the authors propose an "additive model" for AD risk liability. However, there is no formal statistical procedure evaluating the additivity. The forest plots show mean differences, but no formal statistical test shows that the changes are "additive." Moreover, whether the changes are statistically significant (error bars are too large) is unclear. The authors should adjust their claims in the title and throughout the manuscript.

[Minor points (#10 and #11)]

10. Fig 1e. This figure panel is confusing. It is not immediately clear those visual elements represent three loci. There is no reason this needs to be shown in a circular format.

11. The authors present an analysis based on a new AD-related genotyped cohort of 1,559 individuals in the Korean population, increasing the genetic diversity in AD genetics research. The data availability section does not provide accession IDs for the GWAS summary statistics, but the authors did not deposit the single-cell data. Can authors confirm if the analysis results are readily available to researchers in standard repositories?

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have adequately addressed my concerns and have provided several very thorough analyses to support these revisions.

Reviewer #2

(Remarks to the Author)
The manuscript has been significantly improved after the revision, and we are happy to support the publication of this study in Nature Communications. One last minor suggestion would be that due to the small sample size and the limited number of individuals carrying the rare/structural variants, the potential caveat/limitation of the results on the CWAS analysis should be carefully discussed.
I co-reviewed this manuscript with one of my postdoc trainees.

Reviewer #3

(Remarks to the Author)
Kim et al. report a revised manuscript presenting a whole-genome sequencing analysis of Alzheimer's disease (AD) in an East Asian population, analyzing n=1559 individuals from the Korean AD cohort. The revised STR analysis in the APOE region adds values presented by the manuscript—it is great to see those variations are mostly independent from SNVs in the surrounding regions. However, the authors still include bulk of data that do not meet the statistical significance. Overall, the reviewer cannot recommend publishing the current version of the manuscript.

[Major points]

1. The reviewer appreciates the authors' effort to revisit the statistical significance of the reported association. It seems like there are a lot of materials presented in the main figure & text that do not meet the statistical significance after multiple hypothesis corrections. The nominal significant level alone is not sufficient and the reviewer is worried that the results may contain many false positives with the current liberate significance threshold. Similar concerns were expressed by other reviewers.
2. For example, the reviewer had a hard time finding the justification for keeping the suggestive associations kept in the main texts and main display items (for example Table 1). The study focuses on the newly reported association reaching the multiple-hypothesis adjusted significance level at the APCDD1 locus. The results and overinterpretation on other locus should be moved to the supplement.
3. Similarly, the main text in pages 11-12 still speak at length about the SAMD3 region; explanation on that front should be kept briefer.

4. The authors did a phenomenal job answering reviewer #1 regarding the APCDD1 SNP association in other East Asian cohorts. The replication status (association statistics with p-values) should be included in the main texts (either results or discussion) to help the readers contextualize the reported APCDD1 SNP association in the context of published studies.
5. The new extended figure 5 seems to indicate there is no sufficient power for some of the analyses, including CWAS (panel e), CNV associations (panel f), and STR without APOE (panel h). The authors should remove those analyses in the main texts to avoid overinterpretation.
6. CWAS section. The methodology used in the CWAS section still lacks methodological justification and/or references. It seems like the double-dipping is still a potential issue. Permutation test does not fix the double-dipping issue. The Lasso and the burden test cannot be conducted on the same sample.
7. Also, did the authors adjust for covariates in the Lasso regression? The lasso is typically used for predictive models, and post-selection inference with Lasso requires specific statistical procedures (for example, PMID: 25574062).
8. Related to the point above, in the CWAS analysis, 72 categories are first selected for nominal significance, then 13 are identified as belonging to each of intergenic and splice site regions, and these are pass evaluation based on a permutation test. Then, in the next step, all 496 intergenic categories are used for the network analysis. What is the rationale for expanding from the 13 to all 496?
9. In response to reviewer #3–point #5, the authors claim they conducted additional analysis investigating the Samplot, but the data is not provided to the reviewer.
10. The authors add a citation regarding the reviwers concerns about the validity of transferring PRS models between populations. However, the closeness to the threshold of the amyloid-beta PRS (now Extended Figure 15d) is not addressed.
11. The author find the authors have satisfactorily addressed our concerns with the “additive model” of disease by renaming it as “cumulative effects.” They additionally perform a new incremental r^2 analysis; I find this analysis informative, and the reviewer appreciate the effort.

[Minor points]
12. It’s ambiguous what line 117 is saying about the p-values. Is this 2.2e-17 or 2.2 raised to the negative 17th power?
13. Line 140 has two sentences spliced together.
14. Lines 141-142: Please clarify which dataset, Korean WGS or the chip data, the description “523 DAT and 340 CU individuals” refers to.

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
The authors have made substantial improvements in the revised manuscript. The reviewer appreciates the clear formatting of both the response letter and the revised document, which made it very easy to review the updated manuscript.

The reviewer has two minor comments:

1. Abstract. The statement "Moreover, structural variation analysis showed that short tandem repeat expansion was associated with an increased risk of AD" should acknowledge that the CNV association at the HPSE2 locus has borderline statistical significance at the best rather than being strongly associated.

2. Data Availability Section. The reviewer encourages the authors to adhere to open data principles and ensure that resources are readily accessible to the research community. For example, instead of stating that "Korean scRNA datasets are available from the corresponding author," the data should be deposited in a standard repository commonly used in the field. Similarly, the GWAS catalog accession ID for the new GWAS summary statistics should be included in the published version of the manuscript.

These are minor changes, and the reviewer believes the editors can determine whether the authors have sufficiently addressed them—there is no need for further reviewer feedback. Congratulations on bringing such a complex and valuable project to completion.

Manuscript Number: NCOMMS-24-30438A

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

Responses from the Authors for Review Comments:

We appreciate the reviewers' insightful and constructive feedback. We also appreciate the reviewers' comments that our analyses are rigorous and comprehensive and that the manuscript is clear. In this study, we performed comprehensive analyses across various types of genetic variations using whole-genome sequencing data from Korean individuals with and without Alzheimer's disease. We identified associations between genetic variations and AD-related endophenotypes, including amyloid positivity and cognitive testing scores. We also replicated the results of a recent European study that examined short tandem repeats in patients with AD.

We believe that our study has been significantly improved and strengthened by revising the manuscript in response to your comments. Notably, our GWAS summary statistics for the Korean population will be publicly available in the GWAS Catalog upon publication of our study. This resource will be helpful for other researchers in the field to facilitate meta-analyses without sample overlap and to utilize them for various post-GWAS analyses. This will contribute to future research on multi-ancestry studies of Alzheimer's disease.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In this paper by Kim et al, the authors perform a whole genome sequencing analysis of a cohort from Korea with and without Alzheimer's disease. The authors perform comprehensive analyses across multiple forms of genetic variation. The authors also explore the association between genetic variation and AD-related endophenotypes, including amyloid positivity and cognitive testing scores. The authors identify several new statistical associations, though many of these are only nominally significant. The authors also recapitulate the results of a recent study examining short tandem repeats in AD. The analyses are rigorous and comprehensive and overall the manuscript is clear. However, there are some important concerns related to the manuscript.

- In Extended Figures 3A & D, there appear to be a few individuals who were classified as DAT, but have negative AB biomarkers. Can the authors explain this?

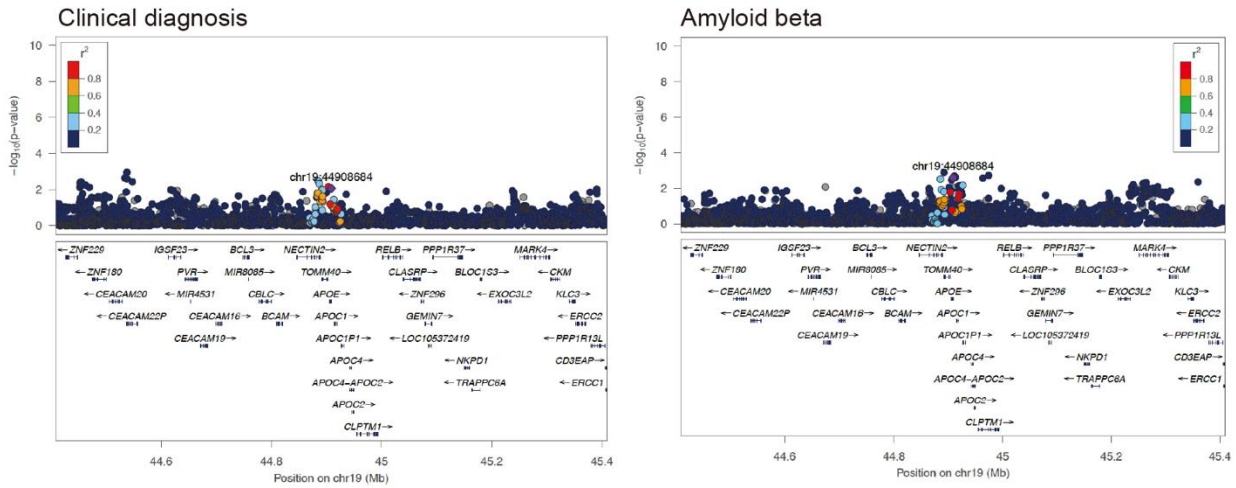
Response: Thank you for your insightful comments regarding the key aspects of defining patient groups. Our study examined individuals with Alzheimer's clinical syndrome from the K-ROAD cohort, categorized as CU, MCI, and DAT based on symptom severity through clinical evaluation, independent of amyloid-beta status. In clinically diagnosed DAT cases, amyloid positivity is observed in approximately 88% of patients, with variations in age and *APOE* status[1]. We have included more details on the sample selection in the Methods section (page 29, lines 549-560).

[**Added to the Methods, page 29, lines 549-560**]

"CU participants were selected based on following criteria: (1) absence of medical history that is likely to affect cognitive function based on Christensen's health screening criteria ... (omitted)"

- For the APOE associations, there is some emerging evidence that there are other alleles near APOE outside of the traditional APOE2/3/4 that may also be associated with AD. It would be interesting if the authors could perform conditional analyses to see if there are additional independent SNP association signals outside of the E2/3/4 alleles.

Response: Thank you for the comments. As recommended, we performed conditional analyses adjusting for the *APOE*2/3/4 genotype by including the rs7412 and rs429358 genotypes as covariates in the single-variant association analysis. After conducting the conditional analysis, we visualized the results using LocusZoom. No significant associations were observed in this region after adjusting for the rs7412 and rs429358 genotypes.



Reviewer-only Figure 1. Regional plots of the *APOE* region for associations after adjusting for APOE status

- For the APCDD1 SNP association, can the authors discuss why this was not uncovered in other AD studies despite the modest size of this study? Is it because the alleles are much more common in the Korean population?

Response: We obtained the allele frequencies (AF) of the *APCDD1* SNP using data from the gnomAD and Korean genome datasets (**Supplementary Table S2a**). The AF of the *APCDD1* SNP is not markedly elevated in East Asian populations. Specifically, the AF was 0.26 in non-Finnish Europeans, 0.17 in East Asians, and 0.15 in the Korean population, including the Korean Genome Project (n=1,832) and the KRGDB cohort (n=2,922). In our Korean whole-genome sequencing (WGS) cohort, the AF was 0.16, and in the Japanese WGS cohort, the was 0.18. Thus, the association observed in our study was less likely to be driven by a significantly higher incidence of AF in the Korean population than in other ethnic groups. However, the consistent signals across three independent East Asian cohorts may suggest a shared genetic signature (Korean WGS, $p = 7.49 \times 10^{-5}$; Korean chip array, $p = 5.85 \times 10^{-3}$; Japanese WGS, $p = 1.79 \times 10^{-3}$). Although we are currently unable to fully explain this observation, we agree with the reviewer that further investigation is essential. As larger East Asian datasets become available, we will continue our efforts to elucidate the underlying causes of this observation and clarify its implications.

[**Added to Supplementary Table S2a**]

Lead variants	EAS	AFR	AMI	AMR	ASJ	FIN	MID	NFE	OTH	SAS
rs28372356	0.17	0.33	0.32	0.26	0.29	0.26	0.30	0.26	0.29	0.38

rs6923619	0.45	0.82	0.33	0.47	0.38	0.34	0.39	0.31	0.44	0.42
rs78009495	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

AF, allele frequency; EAS, East Asian; AFR, African/African-American; AMI, Amish; AMR, Latino/Admixed American; ASJ, Ashkenazi Jewish; FIN, Finnish; MID, Middle Eastern; NFE, Non-Finnish European; OTH, Other; SAS, South Asian.

- For the single variant associations at APCDD1, SAMD3, and PTPRD, the authors should perform statistical fine-mapping to see if they can refine the credible set of potentially causal variants that may aide in the identification of the candidate gene at each of these loci.

Response: As recommended, we have performed statistical fine mapping as part of the colocalization analysis to calculate the 95% credible set. We calculated the 95% credible set, which represented the smallest set of SNPs with a cumulative SNP.PP.H4 greater than 95%. We added a credible set to **Figure 2e**. The full list of SNPs in the credible set is shown in **Supplementary Table S2e**. We have also updated the Results (page 12, lines 219-222) and Methods sections (page 34, lines 732-733).

[Added to Fig. 2e]

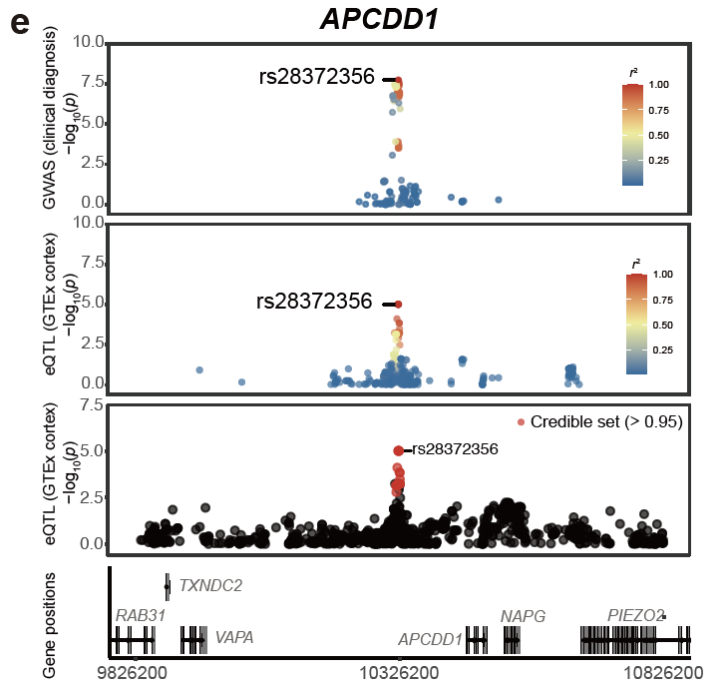


Fig. 2e, Regional plots of the clinical diagnosis GWAS and GTEx cortex eQTLs for *APCDD1* within a 500 kb window of the lead variant (rs28372356).

[Added to the Results, page 12, lines 219-222]

"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, Supplementary Table S2e**)."

[Added to the Methods, page 35, lines 732-733]

"The 95% credible set consisted of the smallest subset of SNPs with a cumulative SNP.PP.H4 of 95%[77]."

- For DCR7 rare variant association, has there been any suggestive signal for this gene in other AD genetics papers? Are the alleles seen in DCR7 in this work present in other populations?

Response: In a previous gene burden association study in Europeans using exome sequencing, we attempted to validate the *DRC7* findings, but did not observe any significant associations[2]. However, when examining the allele frequencies (AF) of rare coding variants of the *DRC7* gene, we noted that out of 16 variants with AF reported in the gnomAD data, 11 exhibited the highest AF in East Asian populations. The higher prevalence of AF in East Asians may explain why significant associations were only observed in this population. We have included this information in **Supplementary Table S2d** and added the relevant details to the **Results** section (page 7, lines 153-155).

Gene name	*p* value	cmac all	Group	Beta	Standard error
DRC7	0.39732141	18.1263486	LOF	0.36903262	0.43396013
DRC7	0.54984239	31.1411525	LOF+REVEL>=75	0.19840916	0.33105019
DRC7	0.80648008	76.2735939	LOF+REVEL>=50	0.05384405	0.21751184
DRC7	0.14519234	427.957392	LOF+REVEL>=25	0.13410228	0.09170974

Reviewer-only Table 1. Summary of gene burden association results for *DRC7* in a European cohort[2]. cmac, cumulative minor allele count.

[Added to the Results, page 7, lines 153-155]

"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 Table S2d**)."

[Added to Supplementary Table S2d]

Variants	EAS	SAS	AFR	AMI	AMR	ASJ	FIN	MID	NFE	OTH
chr16:57726130:CGT:C	0	0	0	0	0	0	0	0	0.0000147	0
chr16:57724783:G:A	0	0	0.00004829	0	0.00006546	0	0	0	0.0000147	0
chr16:57726150:A:T	-	-	-	-	-	-	-	-	-	-
chr16:57702095:C:T	0.0001927	0	0.00002413	0	0	0	0	0	0	0
chr16:57707535:C:T	0.0001922	0	0	0	0	0	0	0	0	0
(omitted)										

EAS: East Asian, SAS: South Asian, AFR: African, AMI: Amish, AMR: Admixed American, ASJ: Ashkenazi Jewish, FIN: Finnish, MID: Middle Eastern, NFE: European (non-Finnish), OTH: others.

- In Figures 2b-c, why did the authors not show the eQTL locus zoom plot using ROSMAP or GTEx data as was used for statistical colocalization (as in Figure 2D)? Similarly, why did they not perform colocalization analyses for the Metabrain QTLs?

Response: In response to the reviewer's suggestion, we conducted co-localization analyses using the MetaBrain eQTL datasets (**Figure 2a, 2b**) (page 11, lines 184-186). These analyses identified significant colocalization between the rs429358 variant and *TRAPPC6A* in the MetaBrain cortex African eQTL dataset. We also generated regional association plots using the ROSMAP and GTEx eQTL datasets, which highlighted the co-localization of two novel lead variants, rs28372356 and rs6923619 (**Figure 2e, Extended Figure 7a**). These plots provide visual confirmation that the GWAS loci align with eQTL signals.

[Added to the Results, page 11, lines 184-186]

"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."

[Added to Fig. 2a,b]

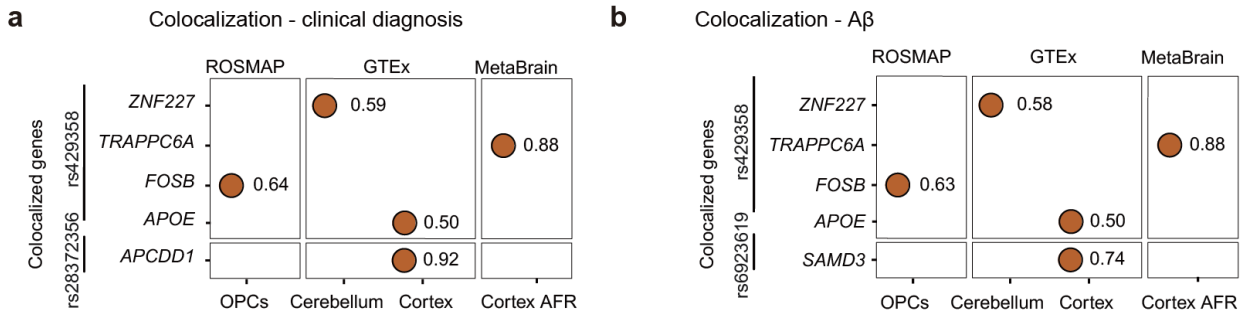


Fig. 2 a,b, Colocalization of GWAS of clinical diagnosis (**a**) and Aβ levels (**b**) with ROSMAP, MetaBrain, and GTEx eQTLs. Colocalizations with posterior probability above 0.5 are shown.

[Added to Fig. 2e]

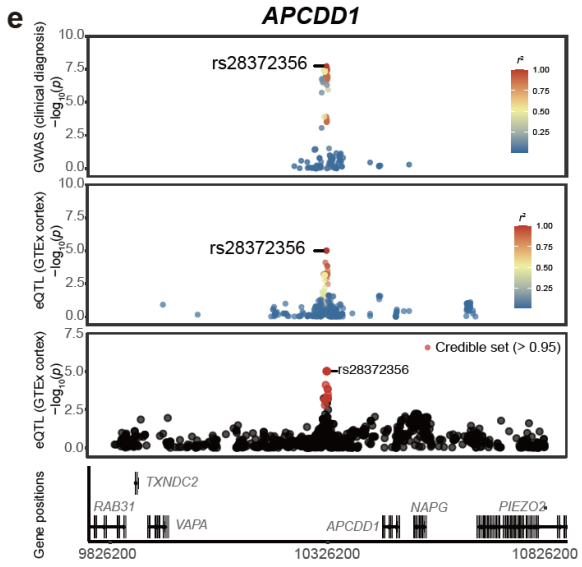
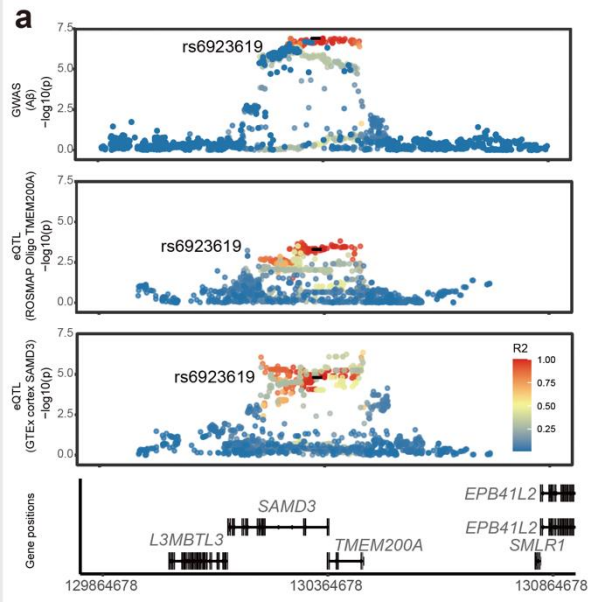


Fig. 2e, Regional plots of the clinical diagnosis GWAS and GTEx cortex eQTLs for *APCDD1* within a 500 kb window of the lead variant (rs28372356).

[Added to Extended Fig. 7a]



Extended Fig. 7a, Regional plot of the Aβ GWAS, ROSMAP oligodendrocyte eQTLs for *TMEM200A*, and GTEx cortex eQTLs (v8) for *SAMD3* within a 500 kb window of the lead variant (rs6923619).

- In the CWAS study, the authors say that the use Lasso regression to prioritize noncoding categories contributing to AB positivity. The authors should better explain how this was done, as I am quite confused by this section.

Response: We apologize for the confusion caused by the insufficient explanation. The primary aim of this analysis was to identify rare noncoding variants associated with AD to utilize these selected annotations in subsequent analyses. For the annotation selection process, we employed LASSO regression.

Initially, we gathered 36 functional and genomic annotations associated with AD in previous studies. The categories derived from these annotations represented sets of variants, with each sample containing a different category (variant). For each sample, we generated data indicating the presence or absence of each category, thus capturing each sample's category profile for use as model input.

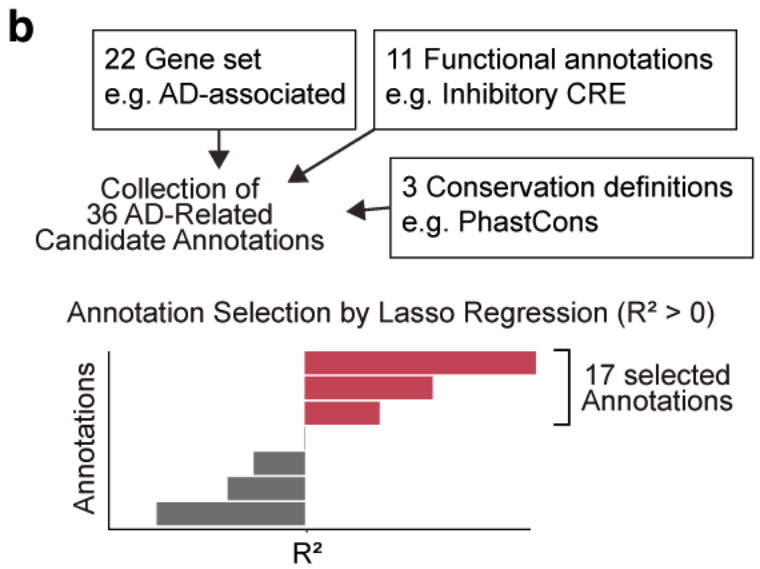
In this dataset, each of the 36 annotations was treated as a separate category group, and these categories were used as features in the LASSO regression model. The model then predicted the phenotype of each sample based on the presence or absence of each category within the sample.

To optimize the LASSO model and enhance its generalizability, we explored 100 lambda values through repeated fivefold cross-validations. We selected lambda, which minimized

the mean cross-validated error, to reduce the overfitting risk. The model performance was assessed using R-squared (R^2) values.

Annotations with positive R^2 values were retained, resulting in the selection of 17 of the original 36 annotations being selected. Only categories with these 17 annotations were used in the subsequent analyses. We have revised **Figure 3b** and expanded the Methods section to enhance the clarity of this approach (pages 36-37, lines 756-770).

[**Modified Fig. 3b**]



[**Added to the Methods, pages 36-37, lines 763-777**]

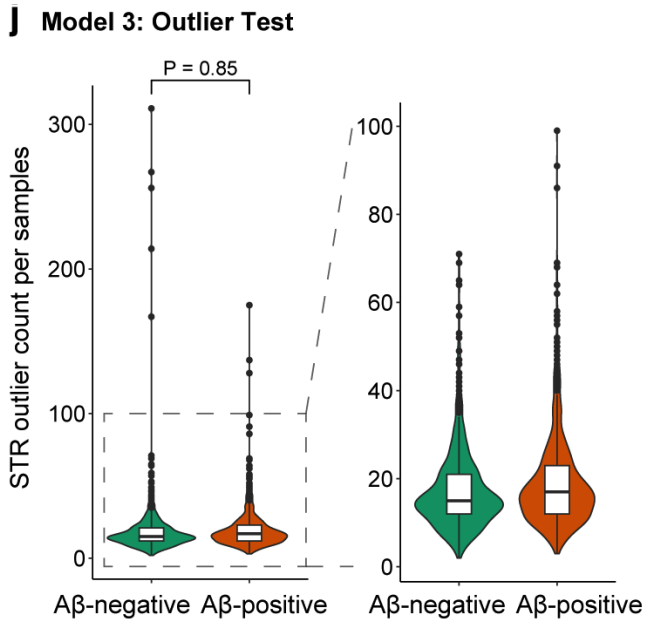
“Annotation selection focused on identifying noncoding categories contributing to Aβ positivity. As mentioned previously, we gathered annotations from various sources, including gene sets, functional annotations, and conservation scores for each variant position, resulting in 36 distinct annotations. These annotations were used to define 29,917 unique categories grouped by the annotation type. Each of the 36 annotations formed its distinct category group, which served as a feature in the LASSO regression model. For each annotation set, we calculated the r^2 values using LASSO regression based on the presence counts of each category in the case and control samples. LASSO regression models were constructed using categories as features and phenotypes (case/control status) as response variables. The Lasso model was tuned with 100 lambda values through repeated five-fold cross-validation, selecting the lambda that minimized the mean cross-validated error to enhance model generalizability and prevent overfitting. The model performance was evaluated using r^2 values. Annotations with positive r^2 values were retained, resulting in the selection of 17 of the original 36 annotations being selected.

Only categories with these 17 annotations were used in subsequent analyses, prioritizing those most strongly associated with Aβ positivity."

- For figure 4j, the authors should provide an additional zoomed in version (e.g., for outlier counts < 100), as it is hard to see the distribution.

Response: Thank you for your comments. An additional zoomed-in version of **Fig. 4j** was added to clearly show the distribution of STR outliers with outlier counts < 100.

[Modified Fig. 4j]



- For the risk score analysis, heterozygous APOE4 should not be considered as "monogenic". I would generally avoid the term "monogenic" here since it implies that all APOE4 carriers have AD, which would obviate the entire analysis.

Response: We agree that the term "monogenic" is inappropriate in this context. Therefore, we have replaced "monogenic" with "*APOE* ε4 genotype" to provide clarity (page 23, lines 410-411).

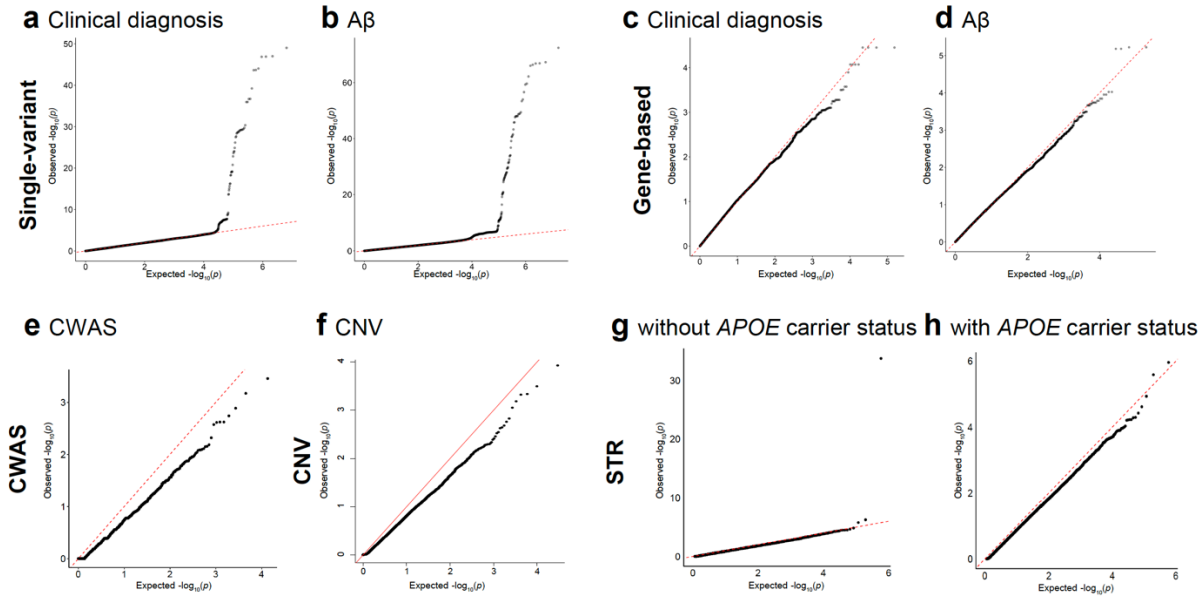
[Added to the Results, page 23, lines 410-411]

"To understand the genetic complexity of AD, we explored phenotypic association of diverse genetic factors across *APOE* ε4 genotype, polygenic, rare noncoding variants, and SVs."

- Throughout, the authors should generate quantile-quantile plots (or suitable alternative) to test for inflation of test statistics. This should be done for at least the GWAS, rare variant analysis, CWAS, and STR analysis under Model 1.

Response: We agree with the reviewer's suggestion and have generated quantile-quantile plots for the GWAS, rare variant analysis, CWAS, CNV, and STR analyses under Model 1. The QQ plots indicate no evidence of inflation. We have included these findings in **Extended Figure 5**.

[Added to Extended Fig. 5]



Extended Fig. 5 QQ plots for GWAS, CWAS, and STR. QQ plots of observed vs. expected −log10(*p*) for GWAS, CWAS, and STR analysis model 1. **a-d**, QQ plots of single-variant association analyses for clinical diagnosis (**a**) and Aβ (**b**), and gene-based association analyses for clinical diagnosis (**c**) and Aβ (**d**). e, QQ plot for CWAS. **f**, QQ plot for CNV. **g**, STR analysis under Model 1 without *APOE* ε4 carrier status. **h**, QQ plot for STR model with *APOE* ε4 carrier status.

QQ plot, quantile-quantile plot; GWAS, Genome-wide association study; CWAS, category-wide association study; CNV, copy number variant; STR, short tandem repeats; Aβ, amyloid beta.

- Throughout, the authors need to more clearly delineate how many hypotheses are being test. For the analyses related to Figure 3g and 3h, these do not appear to corrected for multiple hypothesis testing.

Response: Thank you for your careful review of the critical points. For the analyses related to Fig. 3g-3h, we performed 20 tests consisting of four clinical phenotypes for each of the five risk clusters. Although none of the tests met the threshold for significance after false discovery rate (FDR) correction, our primary goal was to prioritize variants located in intergenic regions that showed the strongest association with the phenotype (increasing AD severity). Based on this criterion, we focused our follow-up analysis on the variants in cluster 73 (C73), which demonstrated the most significant nominal p-values. We have clarified this point in the revised manuscript and Extended Fig. 11 (page 16, lines 285-288; page 38, lines 832-842).

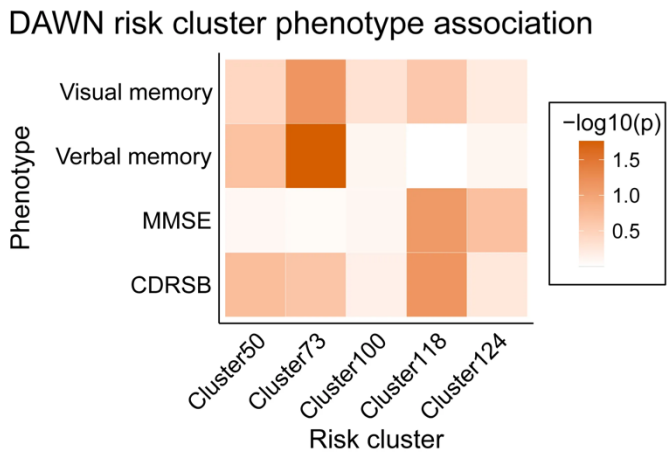
[**Added to the Results, page 16, lines 285-288**]

"To investigate the association between the variants in the five risk clusters and AD, we assessed four AD clinical phenotypes for each cluster (Fig. 3g, Extended Fig. 11). Although none of the 20 tests passed FDR, we prioritized Cluster 73 (C73) for its strongest nominal significance and potential association with increased AD severity."

[**Added to the Methods, page 38, lines 809-819**]

"We conducted a linear regression analysis to evaluate the associations between the five risk clusters (Clusters 50, 73, 100, 118, and 124) and four phenotypic variables (verbal memory, visual memory, CDRSB, and MMSE), resulting in a total of 20 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 20 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 73 (C73), which showed the most significant nominal p-values for follow-up analysis."

[**Added to Extended Fig. 11**]



Extended Fig. 11 Association analysis of AD clinical phenotypes across risk clusters. Evaluation of cognitive function severity across five DAWN-identified risk clusters based on four AD clinical phenotypes (visual memory, verbal memory, MMSE, CDRSB) in Aβ positive samples.
MMSE, Mini-Mental State Examination; CDRSB, Clinical Dementia Rating Sum of Boxes; FDR, False Discovery Rate.

Minor:

- In line 161, it should be TMEM200A, not TEME200A.

Response: We have changed TEME200A to TMEM200A (page 11, lines 200-202).

[**Corrected to the Results, page 11, lines 200-202**]

"Next, for rs6923619 variant, which was associated with Aβ positivity, *SAMD3* was the nearest gene, with significant eQTL signals for *TMEM200A*, *EPB41L2*, *ARHGAP18*, and *SAMD3*."

- In line 201, the authors suggest that they are performing "functional" analyses. I do not think they are performing functional analyses, but rather statistical colocalizations. I would recommend removing the term "functional analyses"

Response: We have changed "functional analyses" to "statistical colocalization" as the reviewer suggested in the Results and Methods sections (page 11, lines 183-184; page 33, lines 669).

[**Revised to the Results, page 11, lines 183-184**]

“To prioritize genes associated with the GWAS loci, we conducted statistical fine-mapping analyses through expression quantitative trait loci (eQTL) colocalization.”

[**Revised to the Methods, page 33, lines 669**]

“Gene prioritization and statistical colocalization analysis”

Reviewer #2 (Remarks to the Author):

In this manuscript, Won and colleagues carried out a GWAS analysis based on the WGS datasets across ~1600 Korean individuals with and without AD. They identified common and rare variants, as well as structural variants (SVs) that are potentially associated with clinical diagnosis of AD and a-beta aggregation. Overall, the genes and loci identified by this study could potentially enhance our understanding of the genetic regulation of AD and help explain the missing variance of AD. However, there are multiple issues with the current manuscript such as its novelty over the existing Asian-based AD GWASs, and the reliability of the genetic association identified, especially regarding the rare variants and SVs. Moreover, the results are generally hard to follow because of the lack of technical/methodological details. The specific comments are as follows:

Major comments:

1. Since there have been multiple studies focusing on the AD GWAS analysis based on Asian populations (https://www.nature.com/articles/s10038-022-01050-z), which showed a comparable or ever larger sample size compared to this manuscript, what are the advantages of the current study? Also, it's important to carry out a comparison with the existing Asian AD GWASs and clarify which loci are reproducible and which are novel.

Response: The current study offers several advantages over previous GWAS on Alzheimer's disease (AD) in Asian populations. Unlike other studies, this study included structural variants (STR and CNV), which have often not been investigated in previous studies. In addition, this study was based on a rich AD phenotype dataset, allowing for a more detailed genetic association analysis. Furthermore, we compared the influence of various variants, including GWAS findings, in **Figure 5** to provide a comprehensive overview of their contribution to AD risk.

For comparison with existing Asian AD GWAS studies, we examined the previously identified loci in our dataset and present this information in a table. In the table below, we summarize the types of genetic variants included, such as common variants, rare variants, noncoding regions, STR, CNV, and various phenotypes analyzed in each study. We also assessed the generalizability of our findings by comparing them with the results of other studies.

References	Common variant (GWAS)	Rare variant (gene burden)	Rare noncoding variant	STR	CNV	Phenotype
K-ROAD (current project)	O	O	O	O	O	clinical diagnosis, Aβ PET imaging, cognitive functions (MMSE, CDR-SB, verbal memory, visual memory)
Shigemizu et al, 2022, Molecular Psychiatry	X	O	X	X	X	clinical diagnosis
Park et al, 2021, Translational Psychiatry	O	O	X	X	X	clinical diagnosis
Jiang et al, 2023, Journal of Neurology, Neurosurgery and Psychiatry	X	O	X	X	X	clinical diagnosis
Shigemizu et al., 2021, Translational Psychiatry	O	X	X	X	X	clinical diagnosis

Reviewer-only Table 2

[**Added to the Introduction, page 4, lines 92-94**]

"Additionally, 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."

[**Added to the Introduction, page 4-5, lines 102-104**]

"We compared our findings with existing Asian AD GWAS to assess the reproducibility and novelty of the identified loci."

[**Added to the Results, page 7, lines 146-147**]

"The three loci (*APCDD1*, *SAMD3*, and *PTPRD*) have not been reported in European or East Asian studies[3,17]."

2. In the rare variant analysis, there are very few individuals (sometimes only 1) carrying the alternative alleles, and therefore I am worried about the reliability of the results. For example, in gene-based association analysis for DRC7, there are many variants where only one individual carries the alternative allele, and therefore might not be convincing to identify their genetic correlation with AD (line 147, Table S2). If the underlying consideration is to build up the association between the gene, rather than individual genetic variants, then the directionality (risk vs protective) of each variant should be considered.

Response: We appreciate your insightful comment regarding the reliability of rare variant analysis. As you pointed out, we examined the directionality of rare coding variants in *DRC7*. Among the 25 variants, 18 showed a higher frequency in the Aβ-negative group, indicating a protective effect, consistent with our gene burden analysis results. We have included this information in the Results section (page 7, lines 153-155) and have updated **Supplementary Table S2d** accordingly.

We aimed to maximize the utility of our whole-genome sequencing data to examine the missing heritability of AD in East Asians, which could not be analyzed using chip array data. Therefore, we incorporated a rare-variant association analysis, as previous studies have indicated that rare variants contribute to missing heritability[3]. We also recognize that a limited number of carriers might have difficulty in convincing the associations of individual rare variants with the disease. Obtaining a statistical significance for individual rare variants requires a much larger sample size than for common variants[4]. For instance, to achieve 80% power for a variant with an minor allele frequency (MAF) of 0.01 and an

OR of 1.4, a sample size of approximately 54,000 would be necessary. Given the current limitations of available cohorts, particularly in East Asian populations, we acknowledge the difficulty of conducting robust single-variant analyses of rare variants.

To overcome this limitation, we conducted a gene burden analysis by grouping variants based on functional units (e.g., genes or categories). This approach allows statistical significance testing with a smaller sample size. While we understand the concerns regarding reliability, we believe that gene burden analysis is the most feasible approach given the current sample size constraints, and publicly providing our results as summary statistics could contribute to meta-analyses with other cohorts to increase statistical power and identify novel associations of rare variants with AD.

[**Added to the Results, page 7, lines 153-155**]

"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 Table S2d**)."

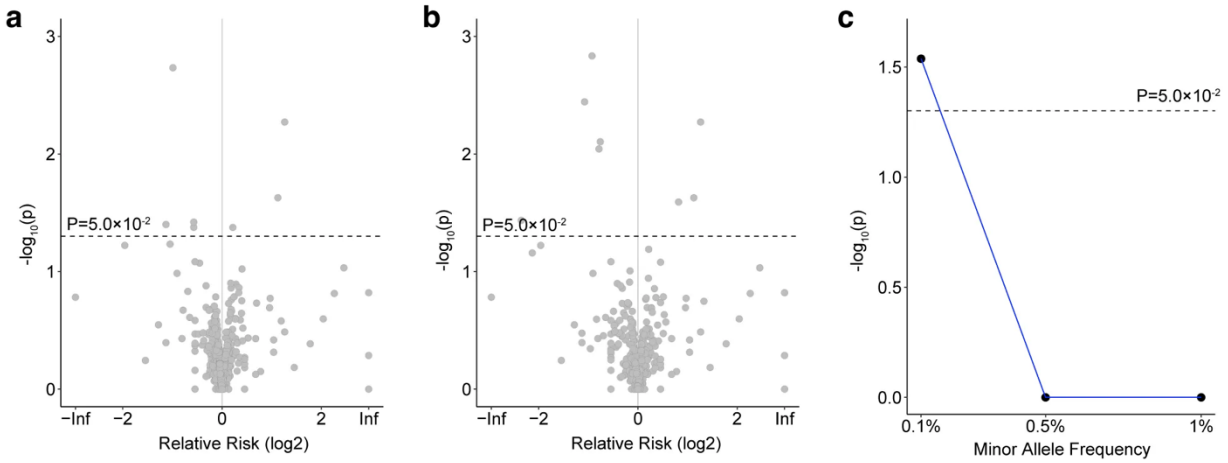
[**Added to Supplementary Table S2d**]

Variants	Number of Aβ-negative carriers	Number of Aβ-postive carriers	Directionality
chr16:57726130:CGT:C	1	0	-
chr16:57724783:G:A	0	0	.
chr16:57726150:A:T	1	0	-
chr16:57702095:C:T	1	1	.
chr16:57707535:C:T	1	0	-
(omitted)			

3. The authors defined common variants using a threshold of MAF>1% while rare variants using MAF<0.1%. Did they exclude the variants with MAF between 0.1% and 1% for analysis? Also, with ~1500 samples, MAF 0.1% means only 1-3 samples with the mutation detected, is it biologically meaningful or statistically feasible to analyze these variants?

Response: Thank you for your important question regarding variant allele frequency. To investigate the variants associated with AD, we conducted a GWAS for common variants with MAF > 1% to identify associations, and a gene burden analysis focused on rare

variants in the coding regions for variants with MAF ≤ 1%. For rare variants in non-coding regions, we employed a Category-Wide Association Study (CWAS)[5,6,7], a method specifically developed using de novo variants that is optimized for handling sparse data and is particularly suitable for analyzing ultra-rare variants. Therefore, we set the MAF threshold to 0.1% to capture the variants present in 1-3 samples, as this condition provided the most optimized results for our CWAS analysis. In accordance with the reviewer's comment, we performed additional analyses using MAF thresholds of 0.5% and 1%, checked the number of categories that passed the nominal significance for each genomic region, and confirmed significance using label-swapped permutation tests. Using this approach, we observed that the significance levels diminished as the variant frequency increased (**Reviewer-only Figure 2c**). To clarify the details of our approach, we have added this content to the Methods section (page 35, lines 730-733).



Reviewer-only Fig. 2

a, CWAS burden test result for MAF 1%. **b**, CWAS burden test results for 0.5% MAF **c**, Relationship between minor allele frequency and statistical significance, illustrating that significance levels diminish as MAF increases. P-values were calculated based on the number of categories that reached nominal significance within each genomic region, which was confirmed using label-swapped permutation tests. Only categories with a sample size of 10 or more were included in this analysis.

[Added to the Methods, page 35-36, lines 735-738]

"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 samples28,29. This optimization guided us to set a MAF threshold of 0.1% to effectively capture these ultra-rare variants."

4. In the introduction, the authors mentioned that “the current GWASs only account for ~15% of the phenotypic variance”, could the authors estimate the additional fraction that this study identified?

Response: Thank you for this insightful suggestion. We agree that estimating the additional fraction of the phenotypic variance explained by the variants identified in this study is informative. To assess this, we employ the incremental R-squared method[8].

We assessed the increase in R^2 by adding the lead variants identified in this study (the K-ROAD study) to previously reported lead variants from a European GWAS (Bellenguez et al.)[9]. Specifically, we conducted logistic regression analyses for each model and calculated the incremental R^2 by determining the difference in Nagelkerke R^2 values between the null model and models that included lead or rare/structural variants.

For clinical diagnosis, the incremental R^2 increased from 12.83% with European loci to 14.68% with the addition of loci identified in the Korean cohort and further to 16.6% with rare and structural variants. For Aβ, the incremental R^2 increased from 7.47% with European loci to 9.83% with Korean loci, reaching 10.7% after including rare and structural variants from the Korean cohort. These findings have been included in **Figure 5a** and details are provided in the Results (page 23, lines 402-409) and Methods sections (pages 42, lines 916-933).

[Added to Fig. 5a]

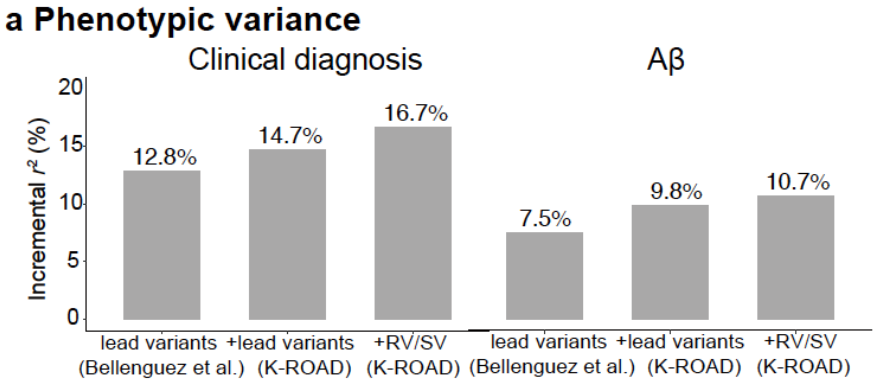


Fig. 5a, Bar plots illustrating the incremental r^2 for clinical diagnosis and Aβ in comparison to previously reported variants[3] (Bellenguez et al.), newly identified variants, RVs, and SVs (K-ROAD).

[Added to the Results, page 23, lines 402-409]

"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. 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 9.83% with the addition of Korean loci and reached 10.7% after the inclusion of RVs and SVs."

[**Added to the Methods, pages 42, lines 916-933**]

"The explanatory power of the identified loci was calculated using incremental r^2 values[102,103]. 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 RVsand SV from the Korean cohort): clinical diagnosis or Aβ ~ rare coding variants (DRC7) + rare noncoding variants (cluster 73) + STR + CNV + model 2."

5. For the rare variants located in the coding regions, the predicted biological effect of these variants on the function/structure of the corresponding protein should be evaluated, such as using PhyloP conservation score, structural domain analysis, and AlphaFold prediction.

Response: We evaluated rare coding variants in the *DRC7* gene using multiple tools, including PhyloP conservation scores, AlphaFold predictions, MetaSVM, SIFT, PolyPhen, and SpliceAI using the Variant Effect Predictor (**Supplementary Table S2d**). Specifically, two variants were predicted to be likely pathogenic by AlphaMissense, four were classified as damaging by MetaSVM, 17 by SIFT, and 13 by PolyPhen. Additionally, 19 variants showed positive PhyloP conservation scores, suggesting evolutionary conservation. We have added a description of the findings to the Results and Methods sections (page 7-8, lines 156-160; page 33, lines 666-668).

[Added to Supplementary Table S2d]

Variants	MetaSVM_pred	SIFT	PolyPhen	am_class	am_pathogenicity	SpliceAI_pred_DS_DL	SpliceAI_pred_SYMBOL	phyloP100way_vertebrate	phyloP17way_primate	phyloP470way_mammalian
chr16:57726130:CGT:C	-	-	-	-	-	0	*DRC7*	-	-	-
chr16:57724783:G:A	tolerated	Deleterious (0.04)	Probably_damaging (0.964)	Likely_benign	0.1836	0	*DRC7*	3.605	0.676	11.693
chr16:57726150:A:T	tolerated	Deleterious (0)	Probably_damaging (0.988)	ambiguous	0.5533	0	*DRC7*	3.955	0.756	11.084
chr16:57702095:C:T	damaging	Deleterious (0.01)	Probably_damaging (0.998)	Likely_benign	0.1152	0	*DRC7*	0.482	-0.271	1.475
chr16:57707535:C:T	damaging	Deleterious (0)	Probably_damaging (1)	Likely_pathogenic	0.7778	0	*DRC7*	2.937	0.589	7.558
(omitted)										

pred, prediction; am_class, The AlphaMissense thresholds are, 'Likely benign' if score < 0.34; 'Likely pathogenic' if score > 0.564; 'ambiguous' otherwise; AG, acceptor gain; DG, donor gain.

[**Added to the Results, page 7-8, lines 156-160**]

"Two variants (chr16:57707535 and chr16:57723121) were predicted to be pathogenic using AlphaMisSense, a machine learning model that assesses missense variant pathogenicity. Additionally, 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."

[**Added to the Methods, page 33, lines 666-668**]

"PhyloP conservation, AlphaMissense, MetaSVM, SIFT, PolyPhen, SpliceAI, and population-specific allele frequencies from the gnomAD database were annotated using the VEP."

6. Figure 2f,g: Why did the authors observe different directionalities of changes for different AD-related variables? Also, can the authors further illustrate the positive vs negative association effects between these two genes and AD severity by analyzing the effect directionality in eQTL vs in AD GWAS?

Response: The CDR score reflects resilience to cognitive decline, demonstrating an inverse directionality compared to AD pathology.

In GWAS data, individuals carrying the rs6923619 variant exhibited lower amyloid-beta levels, and MetaBrain eQTL analysis further demonstrated that the rs6923619 variant was associated with lower expression levels of *SAMD3*. For *SAMD3* differential expression, based on the final cognitive consensus diagnosis, directionality differed between astrocytes and excitatory neurons. Single-cell ATAC-seq analysis identified a peak-to-gene link for rs6923619 specifically at the *SAMD3* locus in astrocytes. Moreover, DEG directionality in astrocytes aligned with both GWAS and eQTL data, indicating a convergent pathway in this cell type (page 12, lines 212-216).

[**Added to the Results, page 12, lines 212-216**]

"Among individuals carrying the rs6923619 variant, lower amyloid beta levels, and *SAMD3* expression were observed in the GWAS and eQTL analyses, respectively (**Table 1, Supplementary Table 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 Table S2h**)."

7. There is a lack of connection between different sections of the results. The author could add more description about the aims of analysis and the logical connection between results in the main text. For examples: Please clarify the analysis of Fig. 2d in the main text (lines 201-207). If the authors only focused on the SAMD3 locus, why were the FOSB and APOE presented in the results? The author should explain why ARHGAP18, PTPRD, EPB41L2, VAPA, and TXNDC2 were included as prioritized genes for functional analysis (lines 227-234).

Response: Thank you for your feedback regarding the coherence of our results. To enhance the clarity of our manuscript, we have implemented several changes to strengthen the connections.

First, we addressed the inconsistencies in the presentation of *FOSB* and *APOE*. We have now provided an explanation of the statistical colocalization findings in **Figures 2a, 2b** and have included all colocalized genes in the Results section (page 11, lines 183-192).

Next, we elaborated on the rationale for prioritizing specific genes for functional analysis. We have focused on the genes prioritized from the significant loci presented in **Figure 2e, 2f** and have included the reasoning for this prioritization to ensure clarity (page 12, lines 218-219).

[Added to Fig. 2a,b]

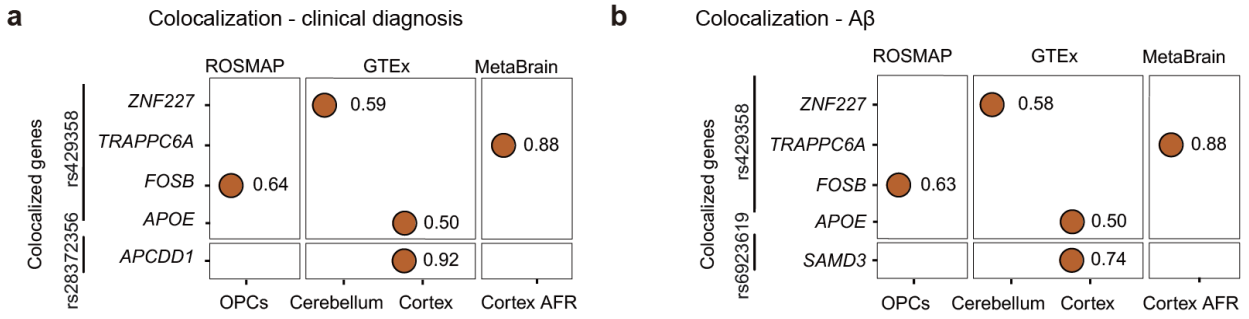


Fig. 2a,b, Colocalization of GWAS of clinical diagnosis (a) and Aβ levels (b) with ROSMAP, MetaBrain, and GTEx eQTLs. Colocalizations with posterior probability above 0.5 are shown.

[Added to the Results, page 11, lines 183-192]

"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, Supplementary Table 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, Supplementary Table S2e**). No significant colocalization was observed for rs78009495."

[Added to Fig. 2e,f]

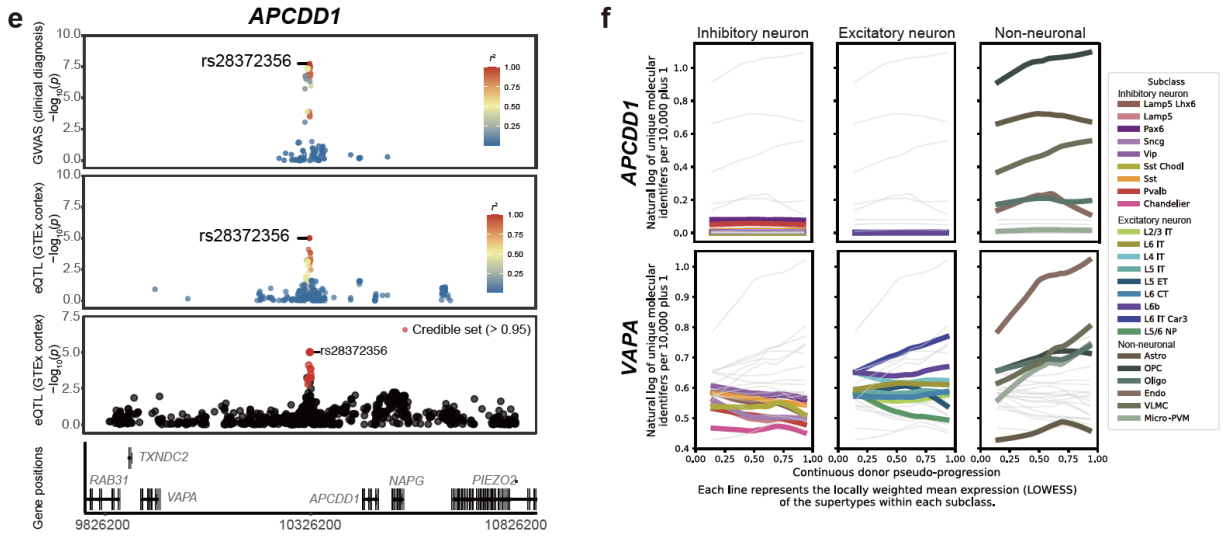


Fig. 2e, Regional plots of the clinical diagnosis GWAS and GTEx cortex eQTLs for *APCDD1* within a 500 kb window of the lead variant (rs28372356). **f**, *APCDD1* and *VAPA* expression across different cell types in response to pseudo-progression from Seattle Alzheimer's Disease Brain Cell Atlas.

[Added to the Results, page 12, lines 218-219]

"We focused on the genes (*APCDD1*, *VAPA*, and *TXNDC2*) prioritized at the newly identified significant locus (rs28372356)."

8. Some Figure/Table legends are incomplete and need to be replenished to increase the readability of the results. There are some examples (but not all of them): In Fig. 1e, what do the different yellow or grey bars mean? In Fig. S1 e-h, please add the legends of bar width.

In Fig. 1f, please add the information on how to build a connection between AD-associated variants and candidate genes into the legend. The author should clarify where the DEGs are from, is the data collected by this study (need to add sample size of each group) or previous studies (need to add reference)? I suggest the author provide a table for details of eQTLs (reference, tissue/cell type, p-value and beta) and DEGs (Fold change, p-value, phenotype, reference/this study). Please add the legends of "variant_type" in Table S2C.

Response: Thank you for your feedback regarding the completeness of our figure and table legends. To enhance the clarity and readability of the results, we have updated the following:

- **Figure 2c** (previously Figure 1e) legend has been updated to include explanatory details regarding the meanings of the colors.
- The legend for **Extended Figure 3e-3h** has been updated to include an explanation of the bar width.
- We have added supplementary tables that detail the eQTLs (**Supplementary Table S2f**) and DEGs (**Supplementary Table S2i**), with an updated legend for "variant_type" (**Supplementary Table S2d**).
- Additionally, we included a citation in the legend indicating that the DEGs were derived from a previous study (page 13, lines 237-238).

[Added to Fig. 2c]

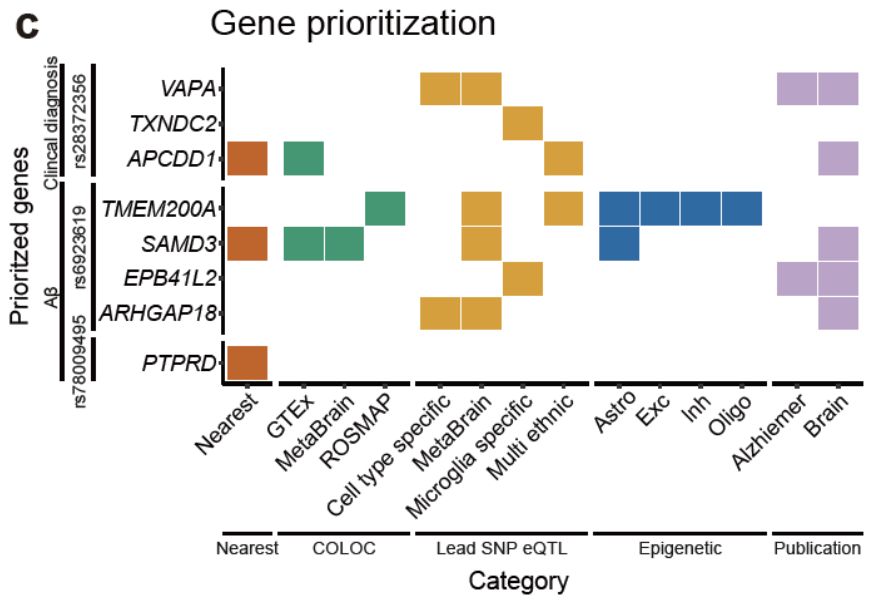
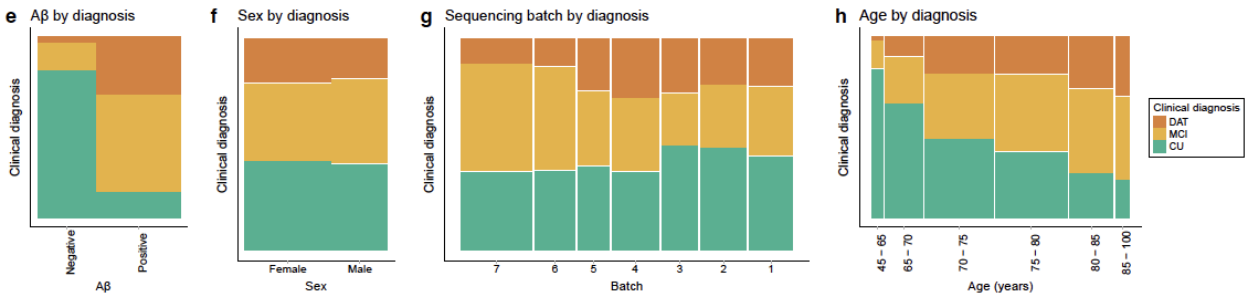


Fig 2c, Heatmap of gene prioritization of loci reaching suggestive significance from single-variant association analysis. The heatmap includes genes that are nearest to lead SNPs (red), eQTL colocalizations (green), eQTL signals associated with lead SNPs (yellow), peak-to-gene connections identified through scATAC (blue), and evidence from prior publications (purple).

[Added to Extended Fig. 3e-h]



Extended Fig. e-h, Distribution of Aβ positivity (**e**), sex (**f**), sequencing batch (**g**), and age (**h**) by clinical diagnosis. The width of each bar corresponds to the sample size for each category.

[Added to the Supplementary Table S2d legend]

"LoF_HC: loss of function high confidence, metaSVM_D: metaSVM deleterious, metaSVM_T: metaSVM tolerance, METASVMandREVEL: both in metaSVM and REVEL"

[Added to Supplementary Table S2f]

Lead variants	Database	Type	Gene	P-value	Beta
rs28372356	Bryios	Endothelial	*VAPA*	0.00030145	-0.282936
rs28372356	Bryios	Excitatory	*LINC01887*	0.00917619	-0.183266
rs28372356	Zeng	Meta	*APCDD1*	0.00119304	na
rs28372356	MetaBrain	cortex-EUR	*VAPA*	0.02757262	0.072526
rs28372356	MiGA	STG_eur	*TXNDC2*	0.04807325	0.273009419
(omitted)					

[Added to Supplementary Table S2i]

Gene	Cluster_id	logFC	logCPM	F	p_val	p_adj.loc	p_adj.glb	coef
ARHGAP18	Inh PVALB CA8 (Chandelier)	0.294325733	7.406645903	43.68052039	2.51E-10	4.47E-08	4.47E-08	cogdx
SAMD3	Exc L2-3 CBLN2 LINC02306	-0.485291151	3.362227449	32.33895601	3.36E-08	1.16E-06	1.16E-06	cogdx
TMEM200A	Exc L2-3 CBLN2 LINC02306	-0.386566835	2.106979682	31.29096153	5.44E-08	1.73E-06	1.73E-06	cogdx
SAMD3	Ast	0.515201234	2.172206563	22.3192623	3.72E-06	6.88E-05	6.88E-05	cogdx
SAMD3	Exc L3-4 RORB CUX2	-0.344308976	3.192376725	20.80518731	7.88E-06	0.000118194	0.000118194	cogdx
(omitted)								

logFC, log fold change; logCPM, log counts per million; p_adj.loc, locally adjusted p-value; p_adj.glb, globally adjusted p-value; coef, coefficient; cogdx, final consensus cognitive diagnosis.

[Added to the Figure legend, page 13, lines 237-238]

"Heatmap of DEGs derived from a previous study[18], according to final cognitive consensus diagnosis across brain cell subtypes."

9. In Fig. 1e-f, peak-to-gene linking was built at cell-type resolution, I was wondering whether this connection of AD-associated loci is consistent with cell-type resolved eQTL. Is there colocalization between AD-associated loci and cell-type resolved eQTLs, and whether these colocalized cell types are consistent with the evidence for the scATAC-seq and DEG analysis?

Response: We performed colocalization, peak-to-gene linking, and DEG analyses to check whether significant results were observed in the same cell types. Although a complete overlap among the cell types was not observed, several notable similarities were identified. For *SAMD3*, peak-to-gene linking was confirmed in astrocytes, which was also significant in the DEG analysis. We have included this information in the Results section (page 12, lines 212-216). Similarly, for *TMEM200A*, oligodendrocytes were significant in the COLOC analysis, which was consistent with the peak-to-gene results.

[**Added to the Results, page 12, lines 212-216**]

"Among individuals carrying the rs6923619 variant, lower amyloid beta levels, and *SAMD3* expression were observed in the GWAS and eQTL analyses, respectively (**Table 1, Supplementary Table 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 Table S2h**)."

10. Did the author consider the effect directionality (risk vs protective) of AD-associated loci for the genetic analysis in Fig. 5? In Fig. 5b, is there a significant difference between the 'APOE only' group and '2/3 factors' groups? In Fig. 5c,d, the author should provide the results of 'APOE only' group. In Fig. 5b-d, why was the PRS calculated using the European GWAS dataset? I would suggest including the PRS group calculated by all the associated loci detected by this study to the analysis in Fig. 5.

Response: We appreciate your feedback regarding **Figure 5**. Our responses to your comments are as follows.

- We evaluated the directionality in **Fig. 5** and found that only rare coding variants in *DRC7* conferred a protective effect. Therefore, we excluded these variants and conducted a reanalysis using rare noncoding variants (page 23, lines 411-413; page 42, lines 937-939).
- We confirmed the significance of differences between the *APOE* ε4 only group and the 2/3 factors groups, and the results have been added to **Extended Figure 15b**. We found a significant difference in Aβ levels and verbal memory between individuals with two factors and those in the *APOE* ε4-only group. We have updated the Results section (page 24, lines 441-443).
- The *APOE* ε4 only results have been indicated in **Fig. 5c and 5d**.
- We selected European GWAS data for PRS calculations because large-scale studies are essential for precise PRS estimation. Currently, the availability of large-scale GWAS focusing on East Asian populations is limited. Previous research [10] has demonstrated the transferability of PRS derived from European GWAS to East Asian populations, underscoring the necessity of employing European GWAS in our analysis. We have clarified this in the Results section (page 23, lines 417-419).
- We also calculated the PRS using the three lead variants identified in our study and included the results in Fig. 5d. We have updated the Results and Methods sections (page 24, lines 455-457; page 41, lines 905-907; page 43, lines 952-953).

[Modified Fig. 5]

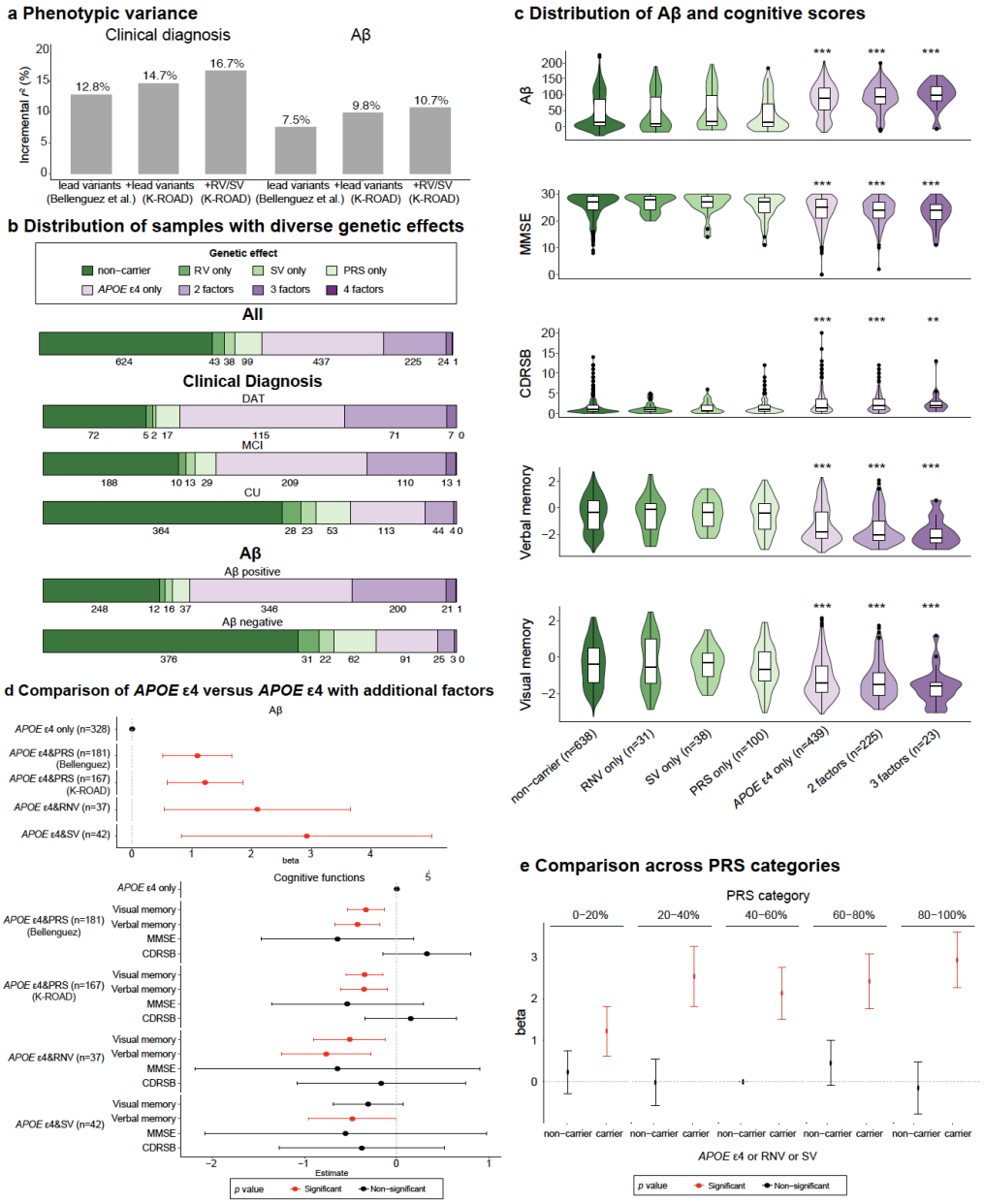


Fig 5. Diverse effects of genetic variants on Aβ levels and cognitive function. a, Bar plots illustrating the incremental r^2 for clinical diagnosis and Aβ in comparison to previously reported variants[3] (Bellenguez et al.), newly identified variants, RVs, and SVs (K-ROAD). **b**, Stacked bar plot depicting 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, MMSE, CDR-SB, verbal memory, and visual memory scores in each group. Wilcoxon rank-sum tests were employed for statistical significance testing, and Bonferroni correction was used to adjust for multiple comparisons. **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 a prior European GWAS[3] (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. Logistic regression was used for Aβ positivity and linear regression was employed for cognitive function for statistical significance testing of the forest plots, adjusting for age, sex, batch, education, and 10 principal components of genetic ancestry. Significance levels are indicated as * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

RV, rare variants; SV, structural variants; DAT, dementia of Alzheimer's type; PRS, polygenic risk score; RNV, rare noncoding variant; SV, structural variant; MMSE, Mini-Mental State Examination; CDR-SB, Clinical Dementia Rating Sum of Boxes; Aβ, Amyloid beta.

[**Added to the Results, page 23, lines 411-413**]

"We excluded rare coding variant carriers, as these variants exhibited a protective effect, opposite to the direction observed for other variants."

[Added to the Methods, page 42, lines 937-939]

"We excluded carriers of rare coding variants, as these variants exhibited a protective effect, which was contrary to the effects observed for other variants."

[**Added to the Results, page 24, lines 441-443**]

"Additionally, there was a significant difference in Aβ levels and verbal memory between the *APOE* ε4-only groups and the two factors (**Extended Fig. 15b**)."

[**Added to the Results, page 23, lines 417-419**]

"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[33]."

[**Added to the Results, page 24, lines 455-457**]

"For this analysis, the PRS was calculated using the three lead SNPs (rs28372356, rs6923619, and rs78009495) identified in this study (K-ROAD)."

[**Added to the Methods, page 41, lines 905-907**]

"Additionally, we calculated PRS using three lead SNPs (rs28372356, rs6923619, and rs78009495) identified in this study. The score function in PLINK was used to compute the PRS using beta values."

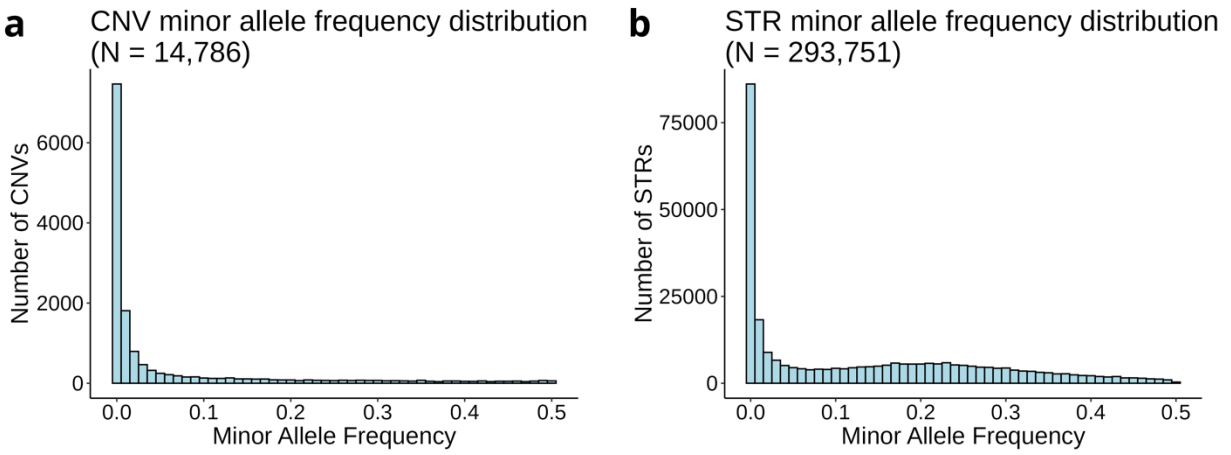
[**Added to the Methods, page 43, lines 952-953**]

"For this analysis, we used the PRS calculated with the three lead SNPs identified in the K-ROAD study."

11. The author should provide minor allele frequency distribution of the structural variants identified, and whether rare SVs were excluded or included in the analysis.

Response: The minor allele frequency for each structural variant (CNV: copy number variant; STR: short tandem repeat) was calculated and its distribution is shown in the figure below. Alleles present in only a single sample were excluded from the CNV analysis. For STR (short tandem repeats), all the alleles were included in the analysis. We have added SV MAF to the Methods section (**Extended Fig. 17**).

[**Added to Extended Fig. 17**]



Extended Fig. 20 Minor allele frequency (MAF) distribution of CNVs and STRs. a, The histogram illustrated the MAF distribution for CNVs across samples. Alleles observed in only a single sample were excluded to ensure a robust distribution profile. **b**, The histogram showed the MAF distribution for STRs. All detected alleles were included in the analysis to capture the full range of STR.

Minor comments:

1. In lines 112-114, is the age of the CU group significantly lower than that of the DAT and MCI groups?

Response: We conducted a t-test to assess age differences and found that the CU group was significantly younger than the MCI and DAT groups (CU vs. MCI, p < 2.2e-16; CU vs. DAT, p < 2.2e-16). To account for this difference, age was adjusted as a covariate in the association analysis. We have included these findings in the Results section (page 6, lines 116-117).

[**Added to the Results, page 6, lines 116-117**]

"CU individuals were significantly younger than both MCI ($p < 2.2^{-16}$) and DAT ($p < 2.2^{-16}$) groups, as assessed by a t-test."

2. In Table 1, could the authors add information regarding the associated phenotypes and the distance between lead variants and TSS of candidate genes?

Response: As suggested by the reviewer, we have added the associated phenotypes and the distance between lead variants and the TSS of the candidate genes to **Table 1**.

The distances were calculated using the UCSC Genome Browser based on representative transcripts from RefSeq and GENCODE. The details of this method are included in the Methods section (page 33, lines 664-666).

[**Added to Table1**]

Gene	Locus	Type	Associated phenotypes	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}
SAMD3	chr6:130364679 (rs6923619)	SNV	Aβ	220,364	0.66 (0.56–0.77)	1.22×10^{-7}
PTPRD	chr9:10825595 (rs78009495)	SNV	Aβ	2,511,349	0.38 (0.26–0.55)	2.07×10^{-7}
DRC7	2 LoF variants, 21 missense variants, 2 splicing variants (See details in Supplementary Table S2d)	Gene-based	Aβ	(See details in Supplementary Table S2d)	0.05 (0.01–0.29)	5.99×10^{-6}
HPSE2	chr10:98736468-98737614	DEL	Aβ	279,391-280,537	1.60 (1.25–1.99)	0.0001

CI, confidence interval; FDR, false discovery rate; SNV, single nucleotide variant; DEL, copy number deletion; Aβ, amyloid beta; TSS, transcription start site.

[Added to the Results, page 33, lines 664-666]

"The distance to the transcript start site was calculated based on representative transcripts from RefSeq and GENCODE in the UCSC Genome Browser."

3. Please unify the description of brain cell types throughout the main text.

Response: We apologize for the inconsistencies in the brain cell-type descriptions. The cell types in the SEA-AD dataset differ from those in the ROSMAP and BTS datasets. To address this, we aligned the major cell types as inhibitory neurons, excitatory neurons, and non-neuronal cells (**Figure 2d, 2f, Extended Figures 6, 7, and 9**). Additionally, we have unified the abbreviations for cell types such as astrocytes (Astro) and oligodendrocytes (Oligo) across the figures. However, owing to the differences in the subclass definitions between the ROSMAP-MIT and SEA-AD datasets, we were unable to fully harmonize the subclass names.

[Modified Fig. 2d]

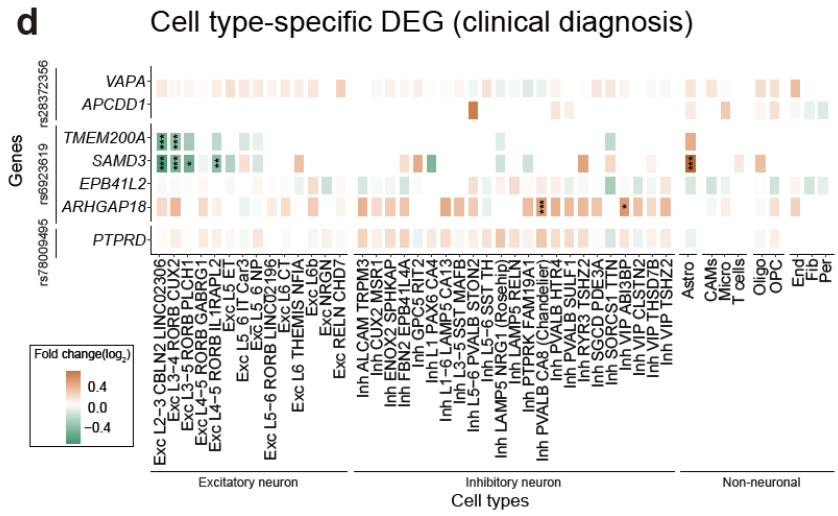


Fig. 2d, Heatmap of DEGs derived from a previous study[18], according to final cognitive consensus diagnosis across brain cell subtypes. Significance levels are indicated as * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

[Modified Fig. 2f]

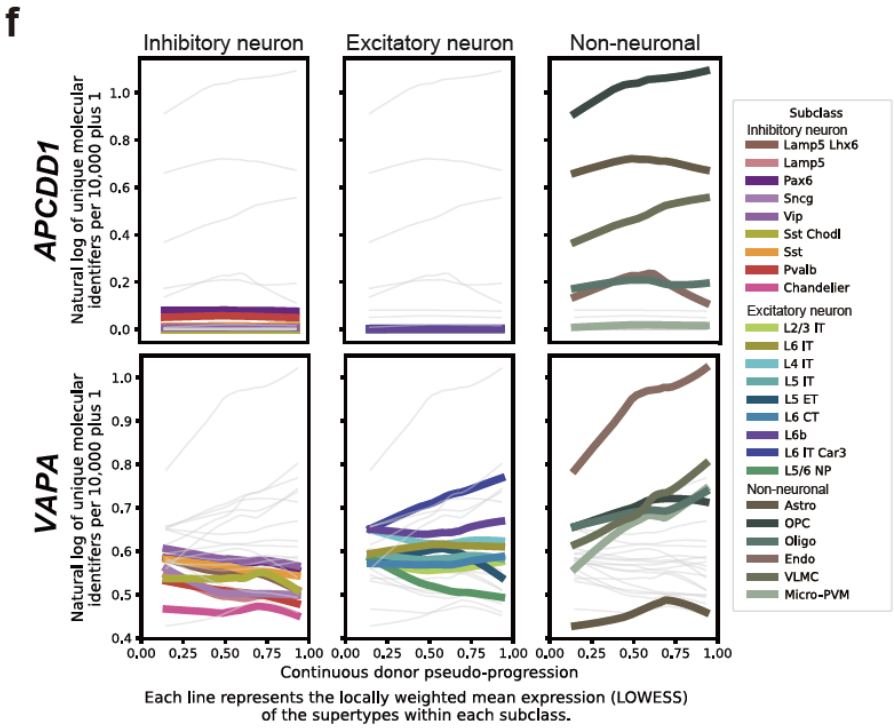
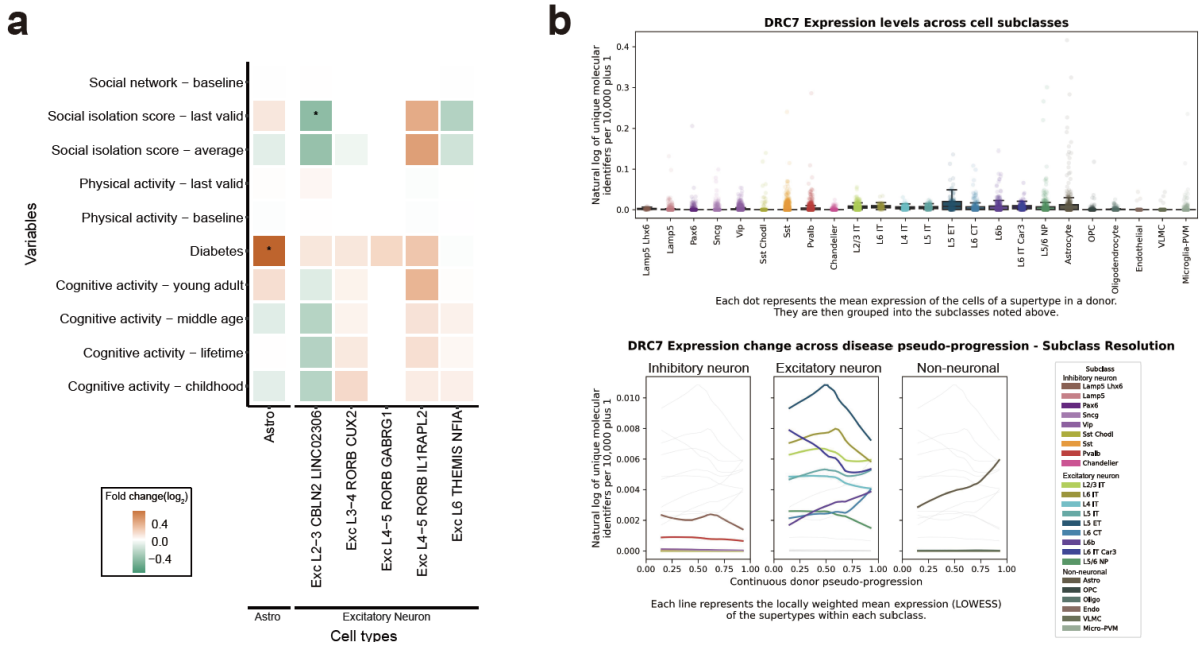


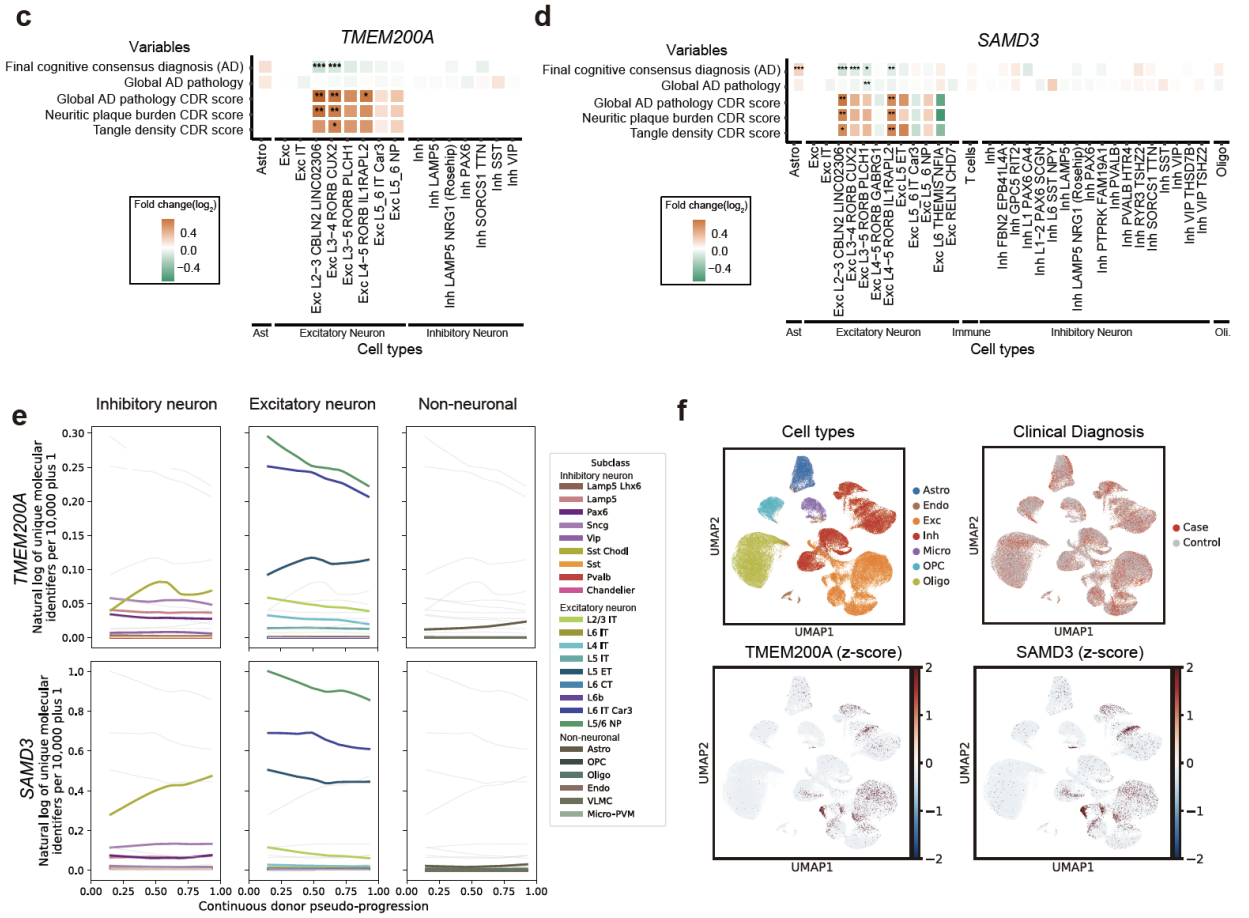
Fig. 2f, *APCDD1* and *VAPA* expression across different cell types in response to pseudo-progression from Seattle Alzheimer's Disease Brain Cell Atlas.

[**Modified Extended Fig. 6**]



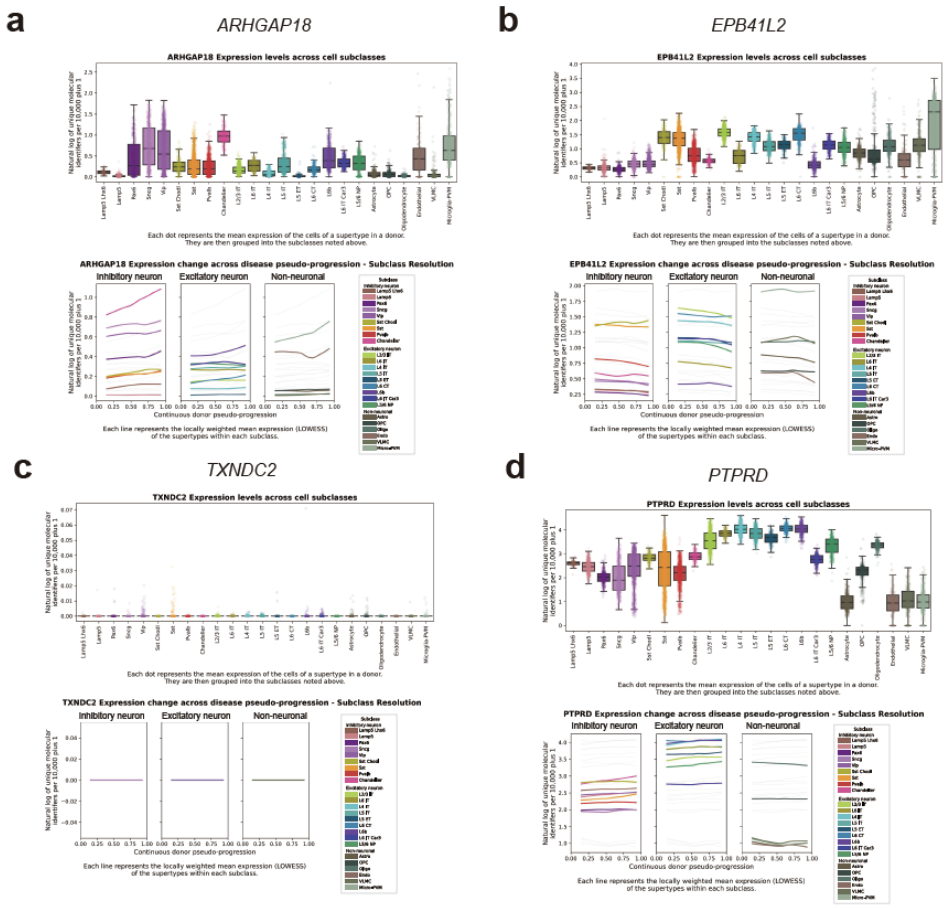
Extended Fig 6, Functional analysis of *DRC7*. **a,** Heatmap of differential expression analysis for *DRC7* across brain sub-cell types from a previous study by Mathys et al. Significance levels are indicated as * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. **b,** Expression levels of *DRC7* across brain sub-cell types and expression patterns across pseudo-progression from Seattle Alzheimer's Disease Brain Cell Atlas. Upper panel: gene expression across brain sub-cell types. Lower panel: expression change pattern with pseudo-progression.

[**Modified Extended Fig. 7c-f**]



Extended Fig. 7 c,d, Heatmap of DEGs for *TMEM200A* (**c**) and *SAMD3* (**d**) across brain cell subtypes. Significance levels are indicated as * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. **e,** *TMEM200A* and *SAMD3* expression across different cell types in response to pseudo-progression from Seattle Alzheimer's Disease Brain Cell Atlas. **f,** UMAP plots using Korean scRNA-seq data, colored based on cell type, clinical diagnosis, and *TMEM200A* and *SAMD3* expression.

[**Modified Extended Fig. 9**]



Extended Fig. 9 Expression of prioritized genes. Expression of *ARHGAP18* (**a**), *EPB41L2* (**b**), *TXNDC2* (**c**), and *PTPRD* (**d**) across brain sub-cell types and expression patterns across pseudo-progression from Seattle Alzheimer's Disease Brain Cell Atlas. Upper panel: gene expression across brain sub-cell types. Lower panel: expression change pattern with pseudo-progression.

OPC, oligodendrocyte precursor cell.

4. In Fig.1, “B” and “C” should be lowercase (Lines 179-180).

Response: We have changed “B” and “C” to lowercase in Figure 1.

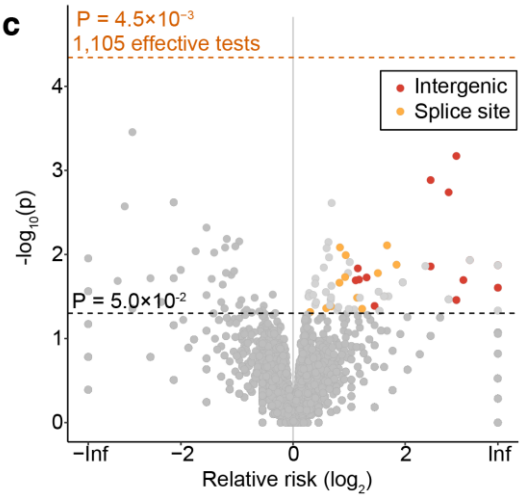
[**Modified the Fig. 1b,c legend**]

"Fig. 1b, 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**)."

5. In Fig. 3c, intergenic loci were marked in red, but the splice site category was not highlighted (line 284).

Response: Following your suggestion, the splice sites have also been highlighted in color for better visibility in **Figure 3c**.

[**Modified Fig. 3c**]



6. The authors may consider performing an enrichment analysis of STRs in different genomic regions in addition to the pie chart (Fig. 4d, lines 354-355).

Response: Thank you for your comments. Instead of comparing the absolute number of STRs, we agree that it would be more appropriate to present the relative distribution through enrichment analysis of STRs across different genomic regions. Therefore, we have replaced **Fig. 4d** with a related illustration that better reflects this approach (page 19, lines 340-342).

[**Modified Fig. 4d**]

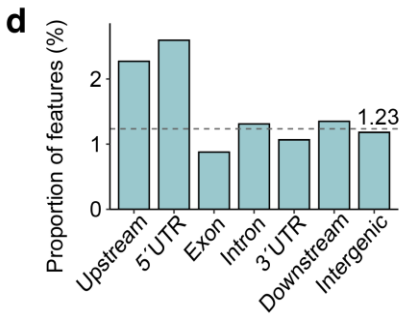


Fig. 4d, Overlap proportion of tandem repeats within each genomic region. The proportion was determined by calculating the length of tandem repeats as a fraction of the total length of the region. The dashed line indicates the average overlap across the entire genome.

[**Modified the Results, page 19, lines 340-342**]

"STRs were more prevalent within 1,000 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**)."

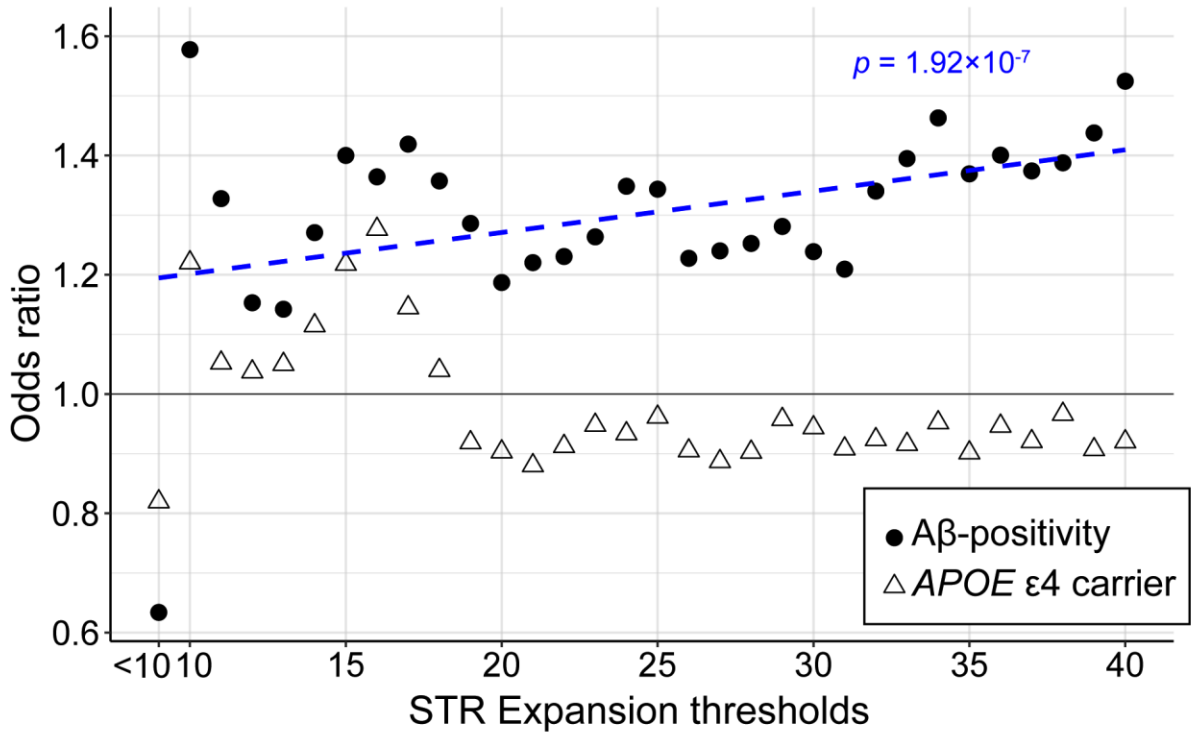
7. Is there a significant correlation between the count of STR outliers and the odd ratio of Aβ-positive per person (Fig. 4k, lines 380-381)?

Response: To assess the correlation between the count of STR outliers and the AD odds ratio, we conducted a linear model analysis by refining the count thresholds and selecting a range from 10 to 40. To account for the influence of *APOE* ε4, we included the odds ratio of *APOE* ε4 carriers as an additional covariate in the linear model. The model used for this analysis was as follows.

$$\text{Aβ-positivity odds ratio} \sim \text{STR}_{\text{outlier count}} + APOE\ \varepsilon 4 \text{ odds ratio}$$

This model showed that the number of STR outliers was associated with an increased AD odds ratio, independent of the *APOE* ε4 odds ratio (p = 9.21e-7). The relevant findings have been added to the extended figure (page 20, lines 364-365).

[Added to Extended Fig. 15]



Extended Fig. 15 Correlation between STR outlier counts and AD odds ratio adjusted for *APOE* ε4 odds ratio. Linear regression model showing the correlation between the number of STR outliers and the odds ratio for Alzheimer's disease (AD), adjusted for the *APOE* ε4 odds ratio. The odds ratio (black circles) increases with the number of STR outliers, independent of *APOE* ε4 odds ratio (white triangles). The dashed blue line represents the regression fit ($p = 9.21\times10^{-7}$).

[Added to the Results, page 20, lines 365-366]

"The odds ratio for Aβ positivity increased with the count threshold (**Fig. 4k, Extended Fig. 15**)."

8. "M" of line 207 should be lowercase.

Response: Thank you for your comment. We changed "M" to lowercase as suggested (page 22, line 395).

9. The length of deletion in the HPSE2 region is inconsistent between Fig. 4b (309kb) and Extended Fig.11d.

Response: Thank you for your comments. For CNVs, mutations within a specific range were considered the same CNV. **Figure 4b** shows the average values, whereas **Extended Figure 11d** shows a variant from a single sample, resulting in different regions in the two figures.

Reviewer #3 (Remarks to the Author):

Kim et al. report a whole-genome sequencing analysis of Alzheimer's disease (AD) in an East Asian population, analyzing n=1559 individuals from the Korean AD cohort. The potential strength of the study is the comprehensive nature of the genetic scan, analyzing single nucleotide variations (SNVs), gene-based rare variant aggregation tests, and copy-number variation analysis. The short tandem repeat outlier analysis (Fig 4k-n) provides new insights into the field. However, the reviewer has concerns about the statistical rigor of the methodology and the interpretation of the results. The manuscript is overall well written. Overall, I cannot recommend publishing the current version of the manuscript.

[Major points (#1-#9)]

1. The reviewer is concerned about the statistical rigor of the analysis, especially the liberating use of a seemingly arbitrary statistical significance threshold. Table 1, which summarizes the "statistically significant" associations reported in the study, lists only one genetic variant (rs28372356). The other reported associations are "suggestive." The authors select some loci with "suggestive" statistical significance and dive into follow-up analysis, for example, in Figure 2. However, using a "suggestive" threshold for prioritizing loci is not a standard practice. Moreover, the authors needed to properly apply multiple hypothesis testing corrections, adjusting for the number of genetic variants and traits considered in their genetic discovery. As such, analyses on "suggestive" loci/associations should be presented in the supplement, not the main text. Authors should provide compelling evidence that their discovery is more than statistical noise or avoid overinterpreting the subsignificant associations.

Response: We fully agree with your concerns regarding the use of a “suggestive” statistical significance threshold. In the initial draft, we focused on the loci with supporting evidence from multiple omics datasets. Although the *SAMD3* region had suggestive significance, it is highlighted in the main figures because of the abundant evidence across multi-omics datasets. However, we agree that the main figures should focus on loci that pass the genome-wide significance thresholds to ensure the robustness of the findings.

In response to your comments, we have thoroughly revised our manuscript and performed additional analyses to concentrate primarily on the significant loci, thereby eliminating overinterpretation of suggestive findings. Specifically, we conducted further follow-up analyses for prioritized genes from the significant loci, *APCDD1* and *VAPA*. We have provided detailed regional eQTL co-localization plots for these genes along with their cell type-specific expression patterns in the main figures (**Figure 2e,f**).

To avoid overinterpretation of the suggestive findings, we moved the previous main figures related to the *SAMD3* region to Supplementary Material (**Extended Figure 7**). We have also reduced the description of these loci in the main text (page 12, lines 217-228).

[Added to Fig. 2e,f]

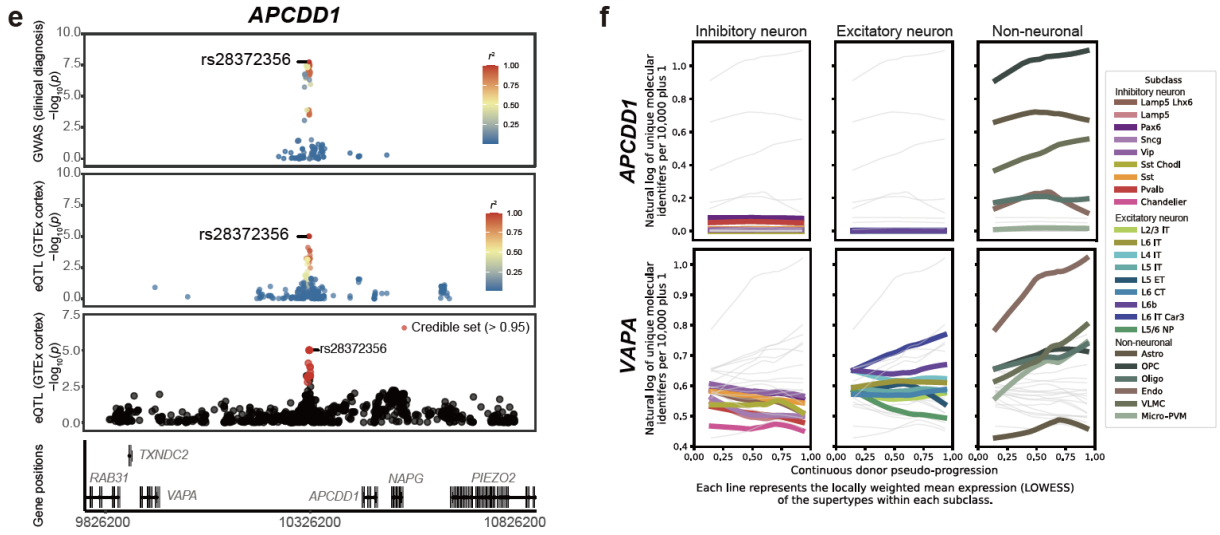
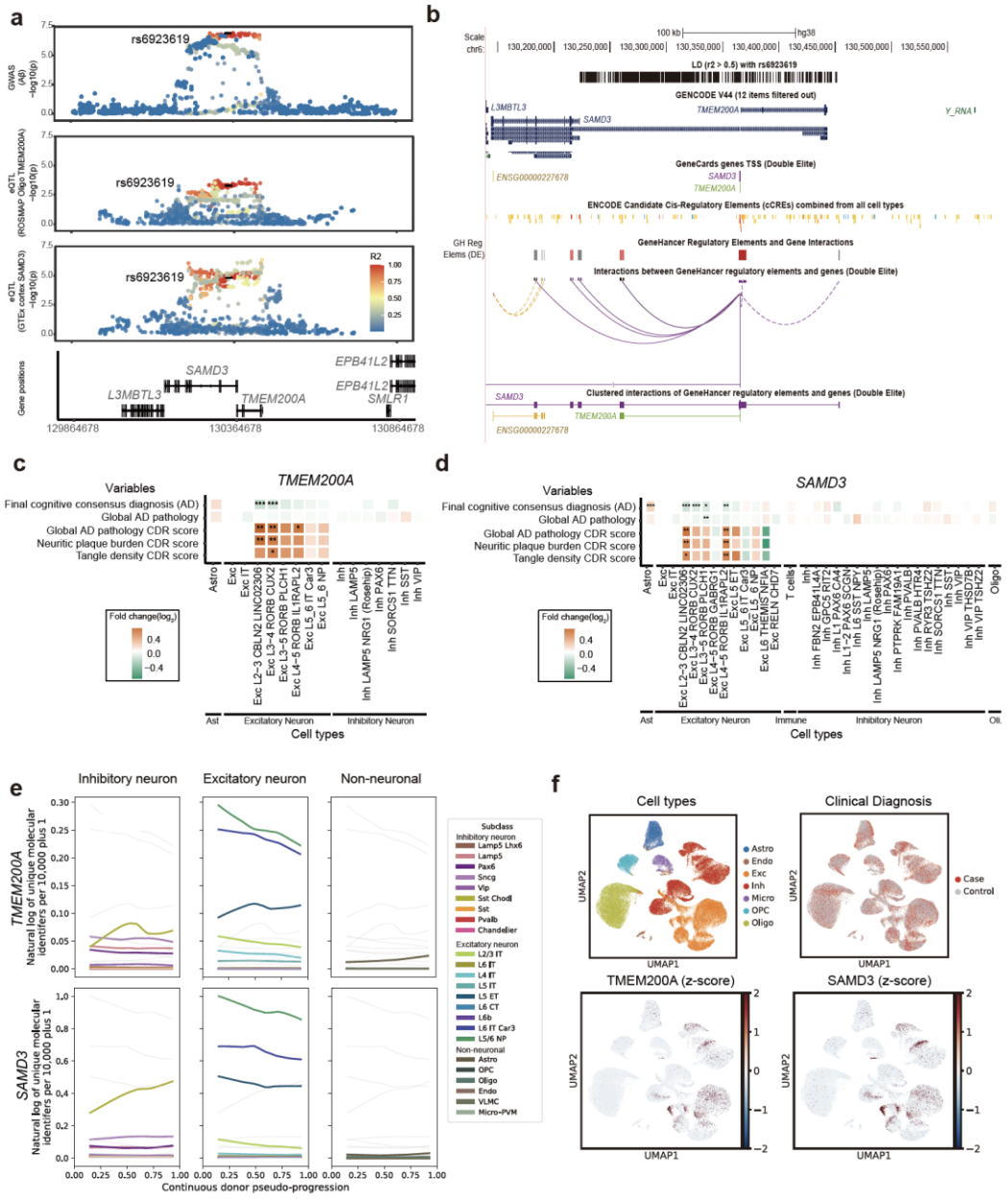


Fig 2e, Regional plots of the clinical diagnosis GWAS and GTEx cortex eQTLs for *APCDD1* within a 500 kb window of the lead variant (rs28372356). **f,** *APCDD1* and *VAPA* expression across different cell types in response to pseudo-progression from Seattle Alzheimer's Disease Brain Cell Atlas.

[Added to Extended Fig. 7]



Extended Fig. 7 Fine-mapping and statistical colocalizations analysis of the *SAMD3* locus. **a,** Regional plot of the Aβ GWAS, ROSMAP oligodendrocyte eQTLs for *TMEM200A*, and GTEx cortex eQTLs (v8) for *SAMD3* within a 500 kb window of the lead variant (rs6923619). **b,** Regional plot showing gene regulatory regions with the lead variant and variants in LD with the lead variant (LD r^2 > 0.5), gene regions from GENCODE, TSS from GeneCards, cCRE regions from ENCODE, and regulatory elements, interactions, and clusters from GeneHencer. **c,d,** Heatmap of DEGs for *TMEM200A* (**c**) and *SAMD3* (**d**) across brain cell subtypes. Significance levels are indicated as * for P < 0.05, ** for P < 0.01, and *** for P < 0.001. **e,** *TMEM200A* and

SAMD3 expression across different cell types in response to pseudo-progression from Seattle Alzheimer's Disease Brain Cell Atlas. **f,** UMAP plots using Korean scRNA-seq data, colored based on cell type, clinical diagnosis, and *TMEM200A* and *SAMD3* expression.

[**Added to the Results, page 12, lines 217-228**]

"To further explore the association between AD and prioritized genes, we examined cell type-specific expression patterns. We focused on the prioritized genes (*APCDD1*, *VAPA*, and *TXNDC2*) at the newly identified significant locus. ... (*omitted*)"

2. Did the authors check if genotyping errors or other technical factors do not drive the genome-wide significant associations (rs28372356)? It would be helpful to report genotype call rate, quality scores, and Hardy–Weinberg Equilibrium Test p-values in the supplement.

Response: As suggested by the reviewer, we checked the genotype call rate, quality score, HWE p-value, and variant quality metrics from GATK (QD, MQ, FS, and SOR) for genome-wide significant associations (rs28372356) and two suggestive associations (rs6923619 and rs78009495). The results are shown in **Supplementary Table S2a**. The lead SNPs exhibited high quality across all quality metrics.

[**Added to Supplementary Table S2a**]

Lead variants	Alleles	Genotype quality	Call rate	HWE p	QD	MQ	FS	SOR
rs28372356	[G,C]	191848.0	0.996568	0.394135	14.85	59.69	0	0.71
rs6923619	[G,T]	678163.0	0.993135	0.980827	22.64	60	0	0.68
rs78009495	[T,C]	60807.7	1	0.665633	15.17	60	0	0.71

HWE_p, Hardy-Weinberg equilibrium p-value; QD, quality by depth; MQ, mapping quality; FS, Fisher's exact test score; SOR, strand odds ratio

3. Did the authors perform power calculations? The fine-mapping analysis for amyloid beta in Fig 2 may not have sufficient power, especially when the lead association does not reach the genome-wide significance threshold.

Response: We appreciate the reviewer's valuable comments. We agree that the loci that do not reach genome-wide significance should not be highlighted in the main figures. We

have moved the original content from **Figure 2** to **Extended Figure 7** and shifted our focus to loci with genome-wide significant associations for follow-up analyses, which are now presented in the main figures (**Figure 2e, 2f**).

[Added to Fig. 2e,f]

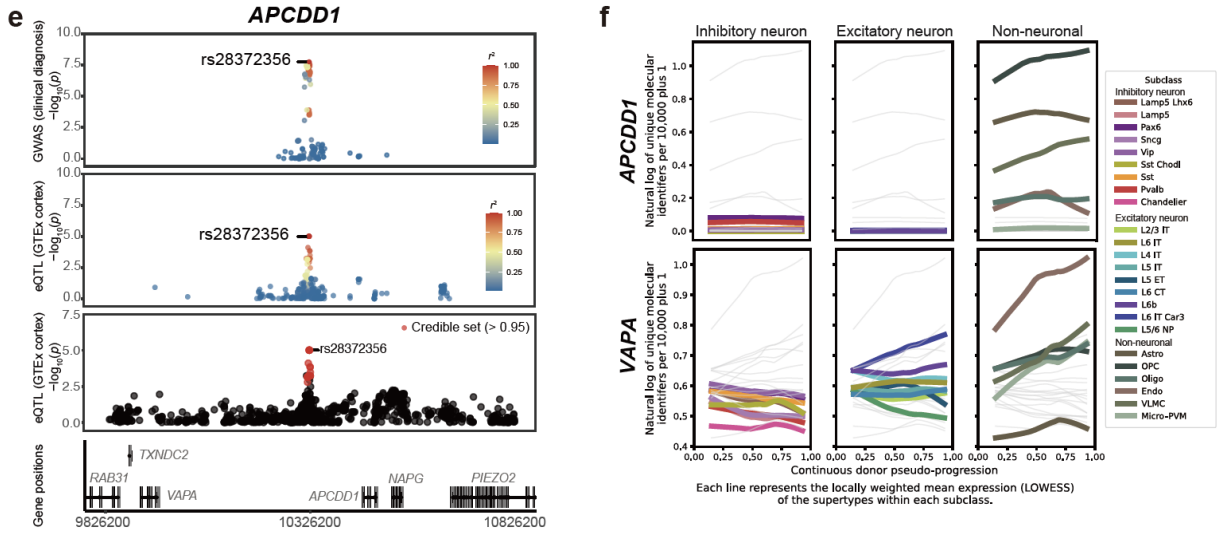
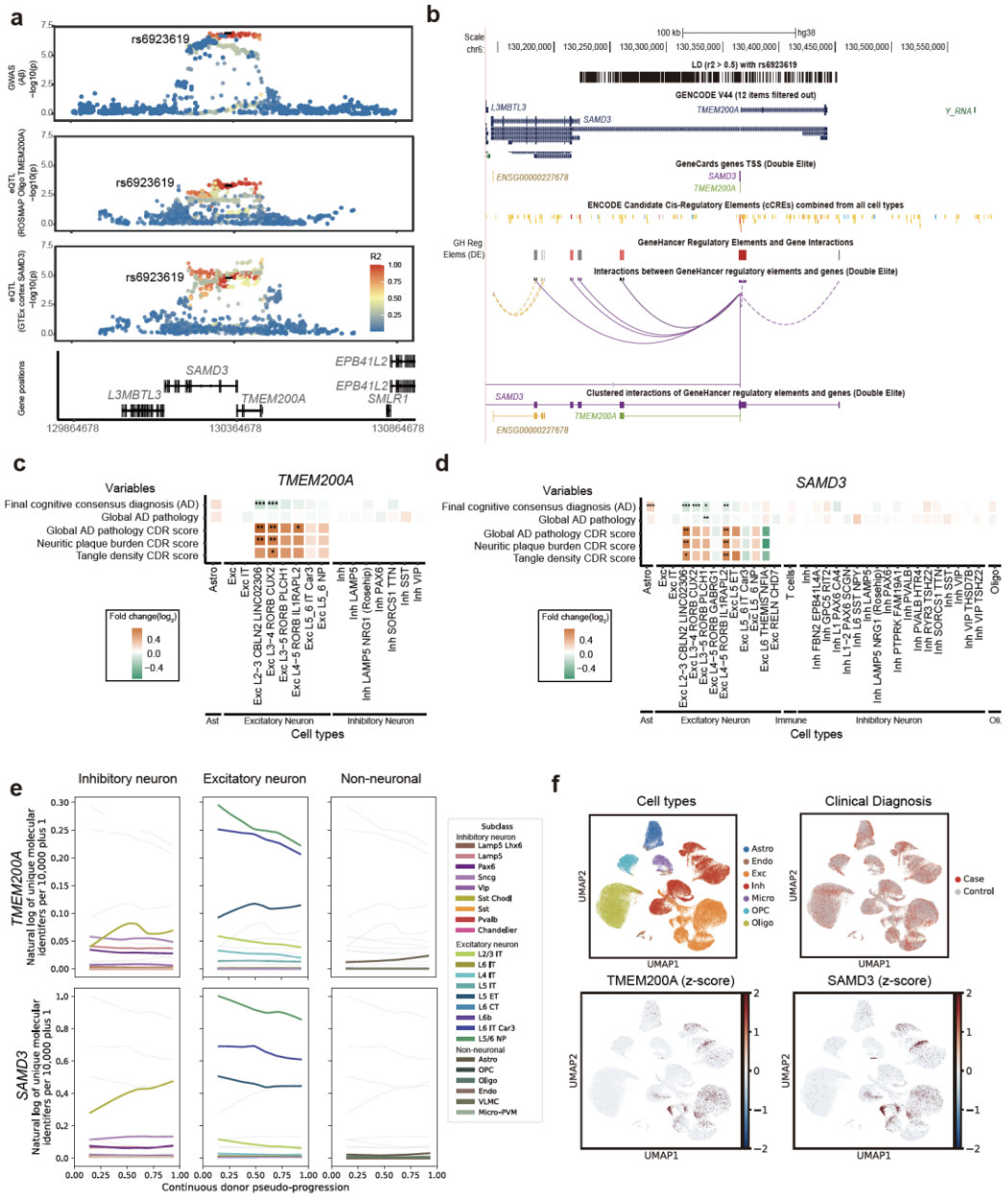


Fig 2e, Regional plots of the clinical diagnosis GWAS and GTEx cortex eQTLs for *APCDD1* within a 500 kb window of the lead variant (rs28372356). **f,** *APCDD1* and *VAPA* expression across different cell types in response to pseudo-progression from Seattle Alzheimer's Disease Brain Cell Atlas.

[Added to Extended Fig. 7]



Extended Fig 7. Fine-mapping and statistical colocalizations analysis of the *SAMD3* locus. **a,** Regional plot of the Aβ GWAS, ROSMAP oligodendrocyte eQTLs for *TMEM200A*, and GTEx cortex eQTLs (v8) for *SAMD3* within a 500 kb window of the lead variant (rs6923619). **b,** Regional plot showing gene regulatory regions with the lead variant and variants in LD with the lead variant (LD r^2 > 0.5), gene regions from GENCODE, TSS from GeneCards, cCRE regions from ENCODE, and regulatory elements, interactions, and clusters from GeneHencer. **c,d,** Heatmap of DEGs for *TMEM200A* (**c**) and *SAMD3* (**d**) across brain cell subtypes. Significance levels are indicated as * for P < 0.05, ** for P < 0.01, and *** for P < 0.001. **e,** *TMEM200A* and

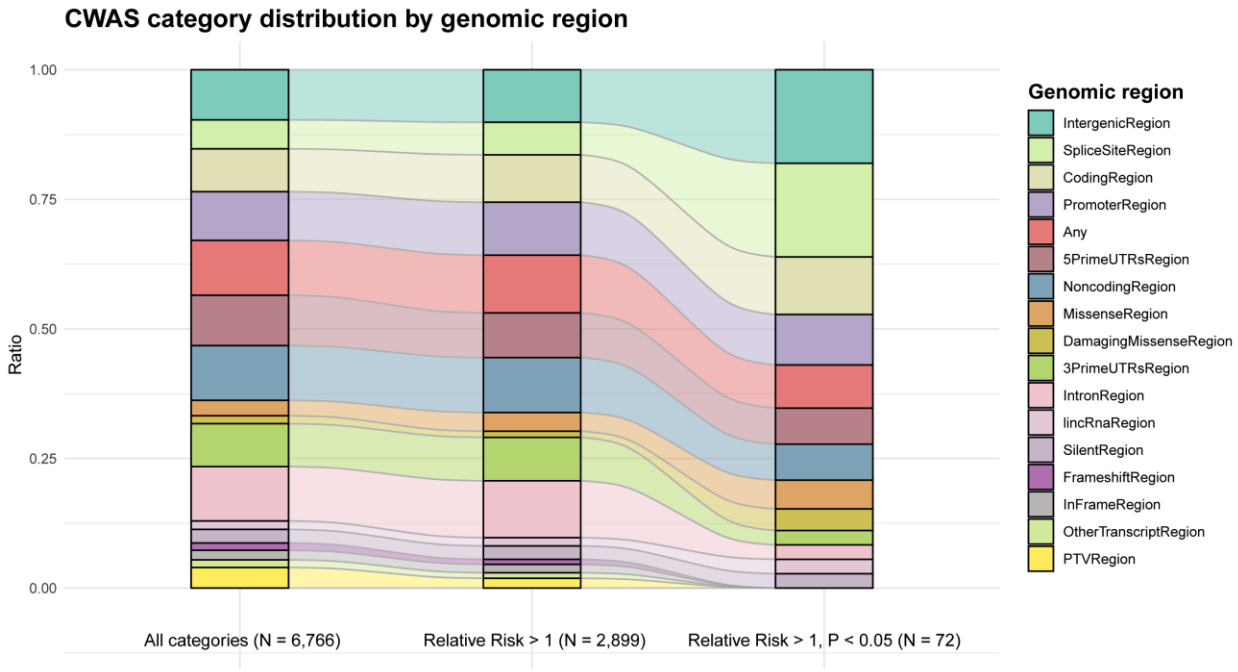
SAMD3 expression across different cell types in response to pseudo-progression from Seattle Alzheimer's Disease Brain Cell Atlas. **f,** UMAP plots using Korean scRNA-seq data, colored based on cell type, clinical diagnosis, and *TMEM200A* and *SAMD3* expression.

4. Similarly, the category-wide association (CWAS) study in Fig 3 does not meet statistical significance (page 16, line 282) for 6766 repeated tests. Moreover, the 6766 categories are selected using the same set of individuals ("double dipping"). The authors should drop overinterpretation (for example, "tend to be enriched" in line 285). The authors should justify the statistical rigor of the CWAS analysis or drop the analysis from the manuscript.

Response: Thank you for your comments. We acknowledge that none of the 6,766 categories passed the threshold for statistical significance after correcting for multiple comparisons. However, in our analysis, we observed a notable pattern among the 72 categories that reached nominal significance, with 26 of these categories belonging to intergenic and splice-site regions (**Reviewer-only Fig. 3**). To validate this pattern further, we conducted a permutation test to assess the significance of the results, as presented in **Figure 3d**. Through this test, we observed significant associations within the intergenic regions, prompting us to perform a network analysis of AD risk factors specific to these intergenic categories.

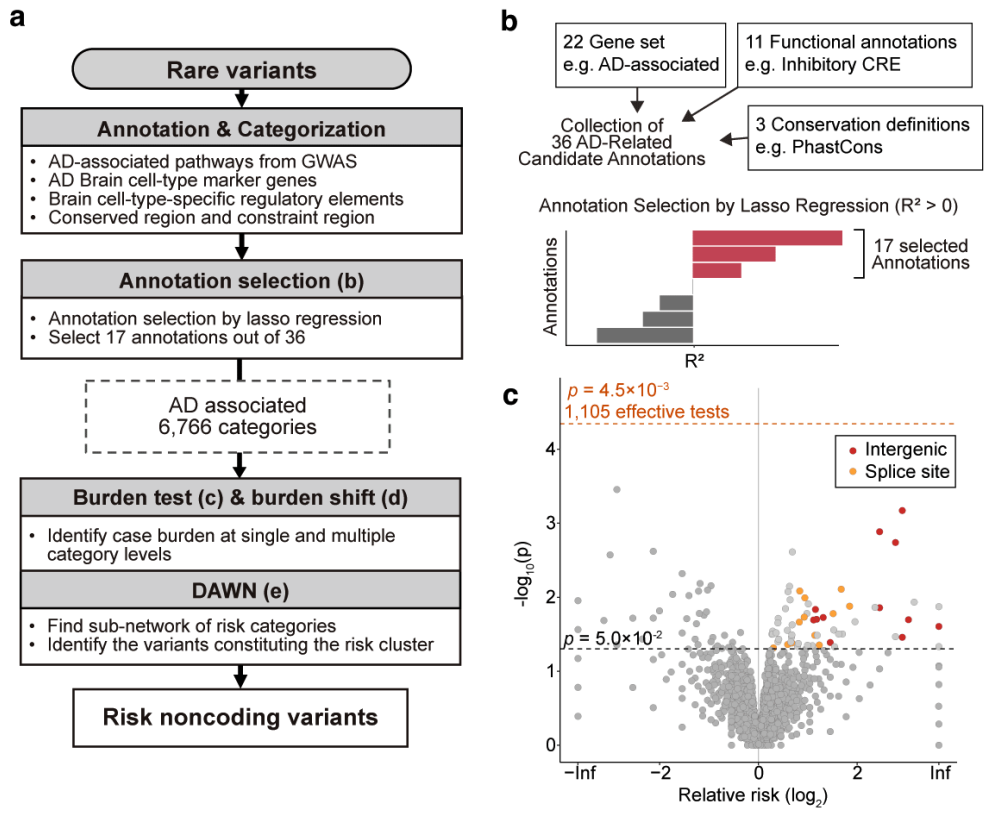
Regarding the concern of double dipping, we refined our analysis to avoid potential overlaps in data use. Initially, our goal was to identify annotations that were highly associated with AD, and then refine the genomic regions within those annotations that exhibited the strongest associations. However, we recognize that this might lead to the risk of double dipping. To address this issue, we revised our approach to focus solely on annotation selection (previously referred to as feature selection) using the LASSO regression. Specifically, we excluded subsequent risk score analyses that utilized the same set of individuals. Consequently, the findings and presentation in **Figure 3b** have been updated to reflect these changes (page 15, lines 263-272).

Finally, we appreciate your suggestion to clarify the language. To avoid overinterpretation, we have revised the relevant expressions (page 15, lines 263-272).



Reviewer-only Fig. 3

[Modified Fig. 3a, b]



[**Added to the Results, page 15, lines 263-272**]

"Because not all annotations effectively indicated AD association, we performed Lasso regression to prioritize annotations contributing to Aβ positivity and selected 17 out of 36 annotations (**Fig. 3b, Supplementary Table S3c**).

Based on these selected 17 annotations, we constructed 6,766 categories and compared their burdens. Although no category showed enrichment after multiple testing corrections (**Supplementary Table 3d**), 72 categories showed nominal significance ($p < 0.05$, relative risk (RR) > 1, binomial test), and intergenic and splice site categories accounted for the highest numbers, with 13 out of 72 categories each (**Fig. 3c**). Permutation tests confirmed statistical significance in the intergenic ($p = 0.029$) and splice site ($p = 0.027$) categories, but not in other noncoding regions (**Fig. 3d, Supplementary Table S3e**)."

5. Figure 4. The cytoband-based Bonferonni correction is not explained in the Methods. The false positive discovery for the ZNF7 locus (line 347) is alarming and casts doubts whether the type I error is properly controlled.

Response: Thank you for your comments. We conducted a visual assessment using Samplot to verify that the results were less likely to be false positives. The QQ plot shows no evidence of inflation, suggesting a low likelihood of false positives (**Extend Figure 5f**).

[**Added to the Discussion, page 28, lines 527-530**]

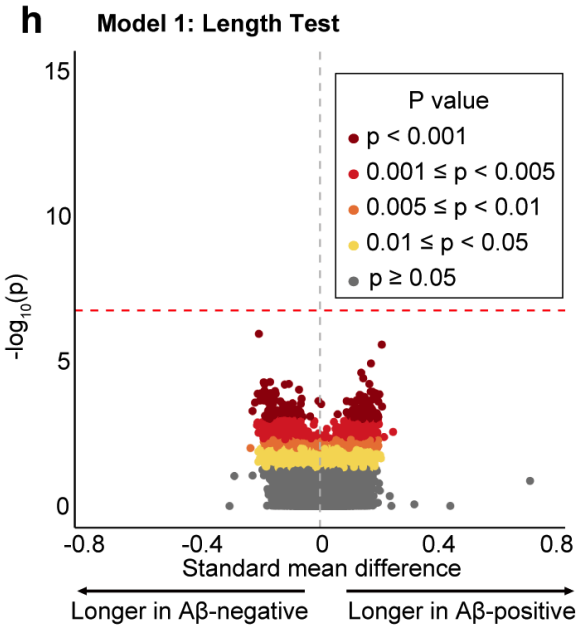
"In addition, cytoband-based Bonferroni correction was applied in the CNV analysis. We also visually verified the presence of CNVs in the identified regions using Samplot and confirmed no evidence of false positives in the QQ plot."

6. Fig 4h. For the Model 1 test for STRs in the APOE locus, the authors should present the association results adjusted for the APOE4 carrier status as the primary results. Does "two additional STRs" meet the multiple hypothesis adjusted threshold?

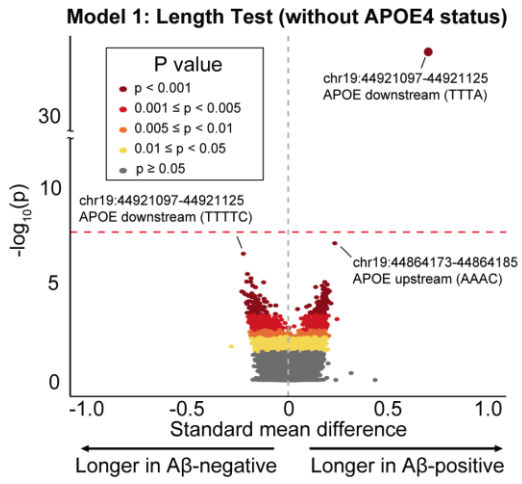
Response: We agree with the reviewer's suggestion that the results adjusted for *APOE* ε4 carrier status should be presented as primary results. Accordingly, we re-analyzed the data considering *APOE* ε4 carrier status as a covariate and replaced **Fig. 4h** with the adjusted results (page 19, line 348-349). In addition, we believe that it is worth including the previous results (without adjustment for *APOE* ε4 carrier status) in the main text because they replicated the findings of a previous European STR study[11]. Therefore, we have included these results in the extended figure (page 19, lines 349-351).

The “two additional STRs” you mentioned did not meet the multiple hypothesis-adjusted significance threshold. Therefore, we have excluded them from the main text to maintain a clear focus on the findings that met the significance threshold.

[Modified Fig. 3h]



[Added to Extended Fig. 14]



Extended Fig. 14. The length of individual STRs was evaluated using a logistic regression model. The volcano plot shows the association between individual STRs (*x*-

axis: difference of the mean length between Aβ-positive and Aβ-negative samples divided by the SD of each STR; *y*-axis: p-value from logistic regression). For this analysis, the *APOE* ε4 carrier status was not included as a covariate because the goal was to test for replicating findings from the European study. The three most significant STRs were located near the *APOE* gene, and the STR that passed Bonferroni correction showed LD with the *APOE* allele (r^2 = 0.56).

[Modified the Results, page 19, lines 348-349]

"Using Model 1 (Length Test) for Aβ positivity, no significant STRs were identified (**Fig. 4h, Supplementary Table S4b**)."

[Modified the Results, page 19, lines 349-351]

"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 = 7.48×10-39; **Extended Fig. 14**). This STR was in LD with the *APOE* ε4 allele (r^2 = 0.56)."

7. Fig 4k-m. The short tandem repeat outlier analysis looks nice. Can authors test whether the identified outliers are in LD with SNVs and if the conditional analysis would attenuate the enrichment observed in Fig 4k-m?

Response: Thank you for pointing out this important aspect of the analysis. We tested whether the STR outliers used in the analysis (n = 15,910) were in linkage disequilibrium (LD) with one significant and two suggestive single nucleotide variants (SNVs) identified through GWAS analysis in this study (rs28372356, rs6923619, rs78009495).

LD calculations were conducted for 84 unique STR outliers located within 2 MB of the three SNVs (**Supplementary Table S4e**). We found that rs6923619 [SNP] and chr6-130282124-CA [STR] were in LD (r^2 = 0.88), whereas the remaining 83 STRs were not in LD with any of the SNVs (r^2 < 0.1). The identified STR outlier was found in one STR out of 1,515 and was present in sample WGS_1393. This sample had 52 STR outliers, and even when this particular STR was excluded from the analysis (reducing the number of STR outliers for WGS_1393 to 51), the results shown in **Figure 4k-4m** remained unchanged. In summary, the results of the STR outlier analysis appeared to be independent of any LD with SNVs.

[**Added to Supplementary Table S4e**]

rsID	SNP	STR	r2

rs6923619	chr6:130364679:G:T		1
rs78009495	chr9:10825595:T:C		1
rs28372356	chr18:10326201:G:C		1
rs6923619	chr6:130364679:G:T	chr6-130282124-130282134-CA	0.882736
rs78009495	chr9:10825595:T:C	chr9-10824378-10824426-TG	0.0337672
rs78009495	chr9:10825595:T:C	chr9-10943762-10943792-AC	0.0211372
rs6923619	chr6:130364679:G:T	chr6-131423268-131423282-AC	0.0046071
rs78009495	chr9:10825595:T:C	chr9-9286873-9286935-AT	0.00424483
rs78009495	chr9:10825595:T:C	chr9-11278563-11278587-AC	0.00379053
rs28372356	chr18:10326201:G:C	chr18-10600552-10600572-TTTGT	0.00322246
rs6923619	chr6:130364679:G:T	chr6-128933230-128933264-TG	0.00257148
rs78009495	chr9:10825595:T:C	chr9-9001973-9001989-TG	0.00245386
rs6923619	chr6:130364679:G:T	chr6-128553155-128553167-AG	0.0024315
rs78009495	chr9:10825595:T:C	chr9-10659662-10659678-TTTAATTT	0.00204239
rs28372356	chr18:10326201:G:C	chr18-9614443-9614455-GCC	0.00178681
rs28372356	chr18:10326201:G:C	chr18-10568642-10568654-AATT	0.00159137
rs78009495	chr9:10825595:T:C	chr9-8962538-8962556-AG	0.00153812
rs28372356	chr18:10326201:G:C	chr18-11529966-11529988-CT	0.00153615
rs78009495	chr9:10825595:T:C	chr9-9211080-9211096-GT	0.00152637
rs78009495	chr9:10825595:T:C	chr9-11832395-11832435-GT	0.00150616
rs28372356	chr18:10326201:G:C	chr18-11462066-11462124-AT	0.00144094
rs78009495	chr9:10825595:T:C	chr9-12807094-12807118-AT	0.00142168
rs6923619	chr6:130364679:G:T	chr6-132153323-132153343-TG	0.00139117
rs78009495	chr9:10825595:T:C	chr9-10393287-10393303-TG	0.00119579
rs6923619	chr6:130364679:G:T	chr6-129619614-129619624-AGGGG	0.00117141
rs28372356	chr18:10326201:G:C	chr18-9341793-9341808-TTTAT	0.00116667
rs6923619	chr6:130364679:G:T	chr6-131692738-131692752-TATATGA	0.00113527
rs6923619	chr6:130364679:G:T	chr6-129051999-129052013-TA	0.00098046
rs28372356	chr18:10326201:G:C	chr18-10998357-10998379-AC	0.00094835
rs28372356	chr18:10326201:G:C	chr18-10589632-10589647-TTG	0.000795712
rs28372356	chr18:10326201:G:C	chr18-11852606-11852618-GTTTTT	0.000720204
rs78009495	chr9:10825595:T:C	chr9-11725915-11725929-AT	0.000714603
rs28372356	chr18:10326201:G:C	chr18-9369136-9369161-AAATA	0.000710295
rs78009495	chr9:10825595:T:C	chr9-11366118-11366138-TA	0.000653575
rs78009495	chr9:10825595:T:C	chr9-8922823-8922839-TTTTTGTA	0.000638103
rs6923619	chr6:130364679:G:T	chr6-128921711-128921723-TTA	0.000596706
rs28372356	chr18:10326201:G:C	chr18-11870553-11870561-AAAT	0.000591488

rs6923619	chr6:130364679:G:T	chr6-128982759-128982774-CCCCA	0.000557701
rs28372356	chr18:10326201:G:C	chr18-11018220-11018244-GT	0.000491117
rs6923619	chr6:130364679:G:T	chr6-129123370-129123388-AT	0.000461488
rs78009495	chr9:10825595:T:C	chr9-11601409-11601441-ATAA	0.000452748
rs6923619	chr6:130364679:G:T	chr6-128577767-128577787-CA	0.0004431
rs28372356	chr18:10326201:G:C	chr18-11538394-11538410-AC	0.000440978
rs28372356	chr18:10326201:G:C	chr18-9149771-9149787-ATTT	0.000407123
rs28372356	chr18:10326201:G:C	chr18-8634269-8634283-TC	0.000391335
rs78009495	chr9:10825595:T:C	chr9-9275833-9275859-TG	0.000390522
rs78009495	chr9:10825595:T:C	chr9-9684693-9684709-TG	0.000382986
rs78009495	chr9:10825595:T:C	chr9-10424313-10424343-TC	0.000351037
rs6923619	chr6:130364679:G:T	chr6-132262805-132262849-ATGG	0.000346669
rs28372356	chr18:10326201:G:C	chr18-8727853-8727873-CT	0.000320264
rs78009495	chr9:10825595:T:C	chr9-10995039-10995053-AC	0.000313723
rs6923619	chr6:130364679:G:T	chr6-132268982-132268990-AAAG	0.000288377
rs78009495	chr9:10825595:T:C	chr9-12125273-12125295-TG	0.000279232
rs28372356	chr18:10326201:G:C	chr18-8427207-8427223-TTTG	0.000275371
rs28372356	chr18:10326201:G:C	chr18-11916147-11916165-TACAAAAAA	0.000268708
rs6923619	chr6:130364679:G:T	chr6-131991009-131991047-AT	0.000243716
rs28372356	chr18:10326201:G:C	chr18-9334505-9334517-GCGG	0.000243095
rs28372356	chr18:10326201:G:C	chr18-11956361-11956379-AT	0.000228725
rs78009495	chr9:10825595:T:C	chr9-8832776-8832794-TC	0.000193465
rs28372356	chr18:10326201:G:C	chr18-12298883-12298895-AAGA	0.000164194
rs78009495	chr9:10825595:T:C	chr9-8934393-8934421-TG	0.000149872
rs78009495	chr9:10825595:T:C	chr9-8948576-8948590-AC	0.000118724
rs6923619	chr6:130364679:G:T	chr6-128758680-128758690-AT	0.000118086
rs28372356	chr18:10326201:G:C	chr18-9542549-9542575-AC	0.00011333
rs28372356	chr18:10326201:G:C	chr18-10731177-10731217-TA	0.000109418
rs28372356	chr18:10326201:G:C	chr18-11354861-11354883-TG	7.95509E-05
rs6923619	chr6:130364679:G:T	chr6-128389423-128389443-TG	7.93098E-05
rs78009495	chr9:10825595:T:C	chr9-10090919-10090959-AC	7.25629E-05
rs28372356	chr18:10326201:G:C	chr18-11962037-11962057-AT	7.16512E-05
rs6923619	chr6:130364679:G:T	chr6-129619345-129619384-AGGGGGAGGGAAA	4.97163E-05
rs6923619	chr6:130364679:G:T	chr6-129718408-129718420-TA	0.000045369
rs6923619	chr6:130364679:G:T	chr6-132299889-132299913-CT	0.000035095
rs28372356	chr18:10326201:G:C	chr18-9876125-9876145-AAAT	2.89056E-05
rs78009495	chr9:10825595:T:C	chr9-10880483-10880503-AAAG	2.85922E-05

rs78009495	chr9:10825595:T:C	chr9-12382165-12382203-TC	2.77366E-05
rs28372356	chr18:10326201:G:C	chr18-11582290-11582302-TA	2.27353E-05
rs28372356	chr18:10326201:G:C	chr18-11845769-11845787-AG	1.96054E-05
rs78009495	chr9:10825595:T:C	chr9-10677134-10677144-AAAAC	1.77718E-05
rs28372356	chr18:10326201:G:C	chr18-11098244-11098274-AC	1.22396E-05
rs6923619	chr6:130364679:G:T	chr6-129847760-129847781-CTTTTTT	9.51741E-06
rs78009495	chr9:10825595:T:C	chr9-10665502-10665528-TA	6.36535E-06
rs6923619	chr6:130364679:G:T	chr6-129381339-129381354-TCTTT	5.09369E-06
rs78009495	chr9:10825595:T:C	chr9-9479347-9479373-CA	4.30602E-06
rs78009495	chr9:10825595:T:C	chr9-10036506-10036534-CA	3.70262E-06
rs78009495	chr9:10825595:T:C	chr9-12087362-12087388-TG	1.82237E-06
rs28372356	chr18:10326201:G:C	chr18-11754998-11755028-TG	3.96765E-07
rs78009495	chr9:10825595:T:C	chr9-12721009-12721031-AA	5.28129E-08

[Add to the Results, page 20, lines 370-371]

"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 Table S4e**)."

[Add to the Methods, page 40-41, lines 882-890]

"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."

8. Polygenic score analysis from European GWAS data. Can authors comment on whether the genetic discovery from European cohorts is readily applicable to the analysis of East Asian samples? Some results (for example, Supplementary Figure 13d) may not survive the statistical significance after adjusting for the number of tests (number of PRS models and tested traits).

Response: We agree with the reviewer's concern regarding the transferability of the PRS from European to East Asian populations. Although it is ideal to calculate the PRS within the same ancestry group, given the limited availability of large-scale AD GWAS studies in East Asian populations, we relied on European GWAS data as an alternative. To address this, we performed cross-ancestry PRS analysis, which has been increasingly applied in recent multi-ancestry studies. Moreover, previous AD studies have demonstrated the transferability of PRS derived from European GWAS to Korean populations[10], and we have cited this reference (page 23, lines 417-419) to support our approach.

[**Added to page 23, lines 417-419**]

"To examine polygenic effects, we employed effect sizes of variants as weights in PRS from large-scale European GWAS data, as prior studies have demonstrated translatability from European to East Asian populations[33]."

9. Based on the results in Fig 5c-d, the authors propose an "additive model" for AD risk liability. However, there is no formal statistical procedure evaluating the additivity. The forest plots show mean differences, but no formal statistical test shows that the changes are "additive." Moreover, whether the changes are statistically significant (error bars are too large) is unclear. The authors should adjust their claims in the title and throughout the manuscript.

Response: As suggested by the reviewer, we have replaced the term "additive model" with "cumulative effects model" to more accurately describe our findings. We also considered alternative terms, such as "combinatorial effects model," if they might further enhance clarity.

[Revised title]

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

[Rephrased in the Abstract, page 3, lines 71-73]

"*APOE* ε4 carriers with high polygenic burden or rare/structural variants exhibited severe cognitive impairment and increased Aβ levels, suggesting a cumulative effects model of AD risk."

[Rephrased in the Introduction, page 5, lines 105-107]

"Our integrative WGS study identified the associations and cumulative effects of various genetic factors with cognitive impairment and Aβ deposition in AD."

[Rephrased in the Discussion, page 27, lines 504-505]

"Our findings suggest a cumulative effects model wherein genetic factors jointly affect susceptibility for AD or Aβ accumulation."

[Rephrased in the Discussion, page 27, lines 508-509]

"However, variations in the individual factors did not cause significant phenotypic changes, indicating their cumulative effects on AD development.

[Rephrased in the Discussion, page 28, lines 532-533]

"Our findings suggest cumulative effects of such genetic variations on AD pathology and support the need for future studies in diverse ancestral populations."

[Minor points (#10 and #11)]

10. Fig 1e. This figure panel is confusing. It is not immediately clear those visual elements represent three loci. There is no reason this needs to be shown in a circular format.

Response: We apologize for the confusion. To enhance clarity, we replaced the circular plot with a heat map (**Figure 2c**), making it more straightforward and easier to interpret.

[**Added to Fig. 2c**]

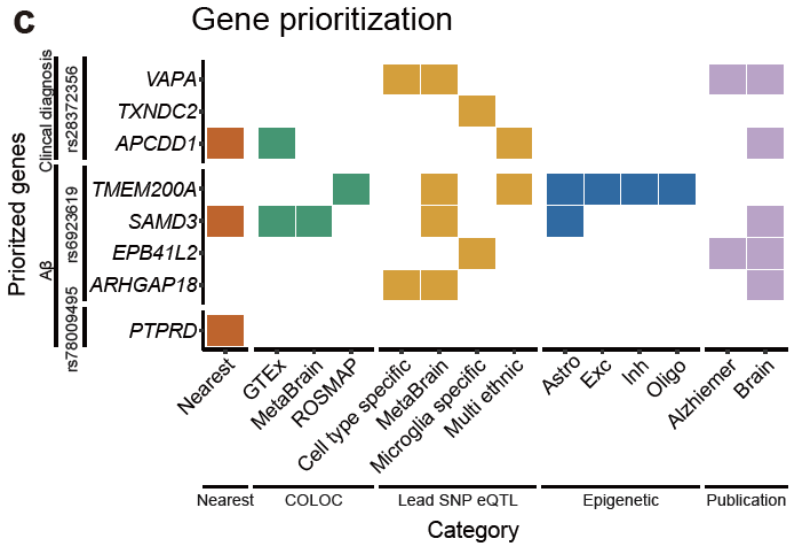


Fig 2c, Heatmap of gene prioritization of loci reaching suggestive significance from single-variant association analysis. The heatmap includes genes that are nearest to lead SNPs (red), eQTL colocalizations (green), eQTL signals associated with lead SNPs (yellow), peak-to-gene connections identified through scATAC (blue), and evidence from prior publications (purple).

11. The authors present an analysis based on a new AD-related genotyped cohort of 1,559 individuals in the Korean population, increasing the genetic diversity in AD genetics research. The data availability section does not provide accession IDs for the GWAS summary statistics, but the authors did not deposit the single-cell data. Can authors confirm if the analysis results are readily available to researchers in standard repositories?

Response: We made the GWAS summary statistics publicly accessible. Upon acceptance of the manuscript, we will upload it to the GWAS Catalog and include the accession IDs in the Data Availability section of the manuscript. Additionally, we plan to share single-cell data following the acceptance of the manuscript.

This resource will be helpful for other researchers in the field to facilitate meta-analyses without sample overlap and to utilize them for various post-GWAS analyses. This will contribute to future research on multi-ancestry studies of Alzheimer's disease.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: We appreciate the reviewers' valuable time and effort in reviewing our manuscript.

Response letter references

1. Ossenkoppele R*, et al.* Prevalence of amyloid PET positivity in dementia syndromes: a meta-analysis. *Jama* **313**, 1939-1949 (2015).

2. Holstege H*, et al.* Exome sequencing identifies rare damaging variants in ATP8B4 and ABCA1 as risk factors for Alzheimer's disease. *Nat Genet* **54**, 1786-1794 (2022).

3. Eichler EE*, et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**, 446-450 (2010).

4. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95**, 5-23 (2014).

5. An JY*, et al.* Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, (2018).

6. Werling DM*, et al.* An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet* **50**, 727-736 (2018).

7. 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**, (2024).

8. 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).

9. Bellenguez C*, et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet* **54**, 412-436 (2022).

10. Jung SH*, et al.* Transferability of Alzheimer Disease Polygenic Risk Score Across Populations and Its Association With Alzheimer Disease-Related Phenotypes. *JAMA Netw Open* **5**, e2247162 (2022).

11. Guo MH, Lee WP, Vardarajan B, Schellenberg GD, Phillips-Cremins J. Polygenic burden of short tandem repeat expansions promote risk for Alzheimer's disease. *medRxiv*, (2023).

Manuscript Number: NCOMMS-24-30438B

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

Responses from the Authors for Review Comments:

We appreciate the reviewers' thoughtful and constructive feedback. In this study, we comprehensively analyzed various genetic variations using whole-genome sequencing data from Korean individuals with and without Alzheimer's disease. Our findings include associations between genetic variations and AD-related endophenotypes, such as amyloid positivity and cognitive testing scores. Additionally, we replicated findings from a recent European study that investigated short tandem repeats in patients with AD.

In response to the reviewers' comments, we revised and refined our description in the manuscript. We believe these revisions have enhanced the quality and clarity of the study. Importantly, we would like to highlight that our GWAS summary statistics for the Korean population will be publicly available in the GWAS Catalog following the publication of this study. We hope this resource facilitates future multi-ancestry studies of Alzheimer's disease by supporting meta-analyses without sample overlap and supports various post-GWAS investigations.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have adequately addressed my concerns and have provided several very thorough analyses to support these revisions.

Response: We appreciate the reviewers' insightful and constructive feedback, addressing which has significantly improved the clarity and robustness of our study.

Reviewer #2 (Remarks to the Author):

The manuscript has been significantly improved after the revision, and we are happy to support the publication of this study in Nature Communications. One last minor suggestion would be that due to the small sample size and the limited number of individuals carrying the rare/structural variants, the potential caveat/limitation of the results on the CWAS analysis should be carefully discussed.
I co-reviewed this manuscript with one of my postdoc trainees.

Response: We appreciate the reviewer's insightful comment. We agree that there are issues due to the small sample size and the limited number of individuals carrying rare/structural variants in the CWAS analysis. In response, we addressed these potential limitations by expanding the discussion in the revised manuscript (page 28, lines 509–517).

[**Added to the Discussion, page 28, lines 509–517**]
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[27]. Future studies should incorporate larger cohorts[34]. Additionally, noncoding regulatory variants exhibit context-dependent effects, thus meriting functional validation, such as through CRISPR-based perturbation[35], to confirm their biological significance.

Reviewer #3 (Remarks to the Author):

Kim et al. report a revised manuscript presenting a whole-genome sequencing analysis of Alzheimer's disease (AD) in an East Asian population, analyzing n=1559 individuals from the Korean AD cohort. The revised STR analysis in the APOE region adds values presented by the manuscript—it is great to see those variations are mostly independent from SNVs in the surrounding regions. However, the authors still include bulk of data that do not meet the statistical significance. Overall, the reviewer cannot recommend publishing the current version of the manuscript.

[Major points (#1-#11)]

1. The reviewer appreciates the authors' effort to revisit the statistical significance of the reported association. It seems like there are a lot of materials presented in the main figure & text that do not meet the statistical significance after multiple hypothesis corrections. The nominal significant level alone is not sufficient and the reviewer is worried that the results may contain many false positives with the current liberate significance threshold. Similar concerns were expressed by other reviewers.

Response: We appreciate the reviewer's thoughtful feedback. We understand the concern regarding the likelihood of false positives with the current significance threshold. As suggested in the major comments, we carefully reviewed and addressed this issue throughout the manuscript. We hope the revisions now improve the clarity and address the reviewer's concerns.

2. For example, the reviewer had a hard time finding the justification for keeping the suggestive associations kept in the main texts and main display items (for example Table 1). The study focuses on the newly reported association reaching the multiple-hypothesis adjusted significance level at the APCDD1 locus. The results and overinterpretation on other locus should be moved to the supplement.

Response: As recommended, we removed suggestive associations from the main display items. Specifically, in Table 1, we removed the suggestive associations (*SAMD3*, *PTPRD*, and *DRC7*). We hope these changes address the reviewer's concerns and improve the focus on the main text.

[**Modified to Table1**]

Table I. Genetic variants significantly associated with clinical diagnosis or Aβ positivity.

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

3. Similarly, the main text in pages 11-12 still speak at length about the SAMD3 region; explanation on that front should be kept briefer.

Response: We agree with the reviewer's suggestion. As suggested, we revised the explanation of the *SAMD3* region and *PTPRD* for conciseness (page 11, lines 200–203). Previously, we provided detailed explanations on how genes were prioritized in these regions, but now we simplified this section by only listing the prioritized genes. Additionally, we removed the mention of *SAMD3*, *PTPRD*, and *DRC7* from the Abstract (page 3, lines 65–66).

[**Modified to the Results, page 11, lines 200–203**]

Next, for rs6923619 variant, associated with Aβ positivity, *TMEM200A*, *EPB41L2*, *ARHGAP18*, and *SAMD3* were identified as prioritized genes[23-25] (**Extended Fig. 7**). Finally, for rs78009495, associated with Aβ positivity, the *PTPRD* gene was prioritized[26].

[**Modified to the Abstract, page 3, lines 65–66**]

Our GWAS analysis identified a previously unreported locus for common variants (*APCDD1*) associated with AD.

4. The authors did a phenomenal job answering reviewer #1 regarding the APCDD1 SNP association in other East Asian cohorts. The replication status (association statistics with p-values) should be included in the main texts (either results or discussion) to help the readers contextualize the reported APCDD1 SNP association in the context of published studies.

Response: We appreciate the reviewer's suggestion and are pleased that our efforts to address the *APCDD1* SNP association in other East Asian cohorts were well received. In response, we incorporated the replication statistics into the main text for clarity and coherence (page 7, lines 142–144).

[**Modified to the Results, page 7, lines 142–144**]
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, $p = 1.79 \times 10^{-3}$).

5. The new extended figure 5 seems to indicate there is no sufficient power for some of the analyses, including CWAS (panel e), CNV associations (panel f), and STR without APOE (panel h). The authors should remove those analyses in the main texts to avoid overinterpretation.

Response: Thank you for your feedback. We acknowledge the reviewer's concern regarding insufficient statistical power for the CWAS, CNV, and STR analyses. In line with the comment, we revised the Results (page 19, lines 321–323) and Discussion (page 28, lines 519–522) to avoid overinterpretation in the CNV section. We revised the Abstract to remove mentions of CNVs (page 3, lines 69–70).

The QQ plot in the context of CWAS corresponds to burden tests on a single category of rare noncoding variants. These variants within a single category are extremely rare (minor allele frequency < 0.1%), and their effect sizes are smaller compared to coding variants. Genome-wide association testing for rare noncoding variants results in an extensive multiple-testing burden because of several hypotheses. Therefore, after we checked that the intergenic variants showed the highest burden with nominal significance in single category-level tests, we examined the statistical significance by permutation testing in multiple category-level tests, following previously described methods[1,2].

In the case of STR analysis, the QQ plot in our extended figure represented Model 1, which did not yield a significant result. However, we observed significant associations with Model 3. The reason for including the results of Model 1 in the main text is that they align with the findings of the European STR study[3], making it necessary to include them.

We acknowledge the reviewer's concern regarding the insufficient statistical power suggested by the QQ plot for CNVs. However, we believe it is important to present these findings with appropriate interpretation. In our study, no CNVs passed the Bonferroni correction. However, the *HPSE2* gene exhibited the most significant signal in the cytoband-based analysis ($p = 0.036$), and its presence in AD samples was visually confirmed (Extended Fig. 12). Given the potential relevance of this CNV, we included this analysis in the main text alongside the two structural variant analyses (CNVs and STRs). For CNV analysis, we acknowledge the limitation of statistical power and aim to emphasize the need for further validation in larger datasets or independent cohorts. This approach aligns with previous WGS studies on AD[4,5]. Acknowledging the reviewer's

concern regarding potential overinterpretation, we modified the Results (page 19, lines 321–323) and Discussion (page 28, lines 519–522).

[**Modified Results, page 19, lines 321–323**]
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).

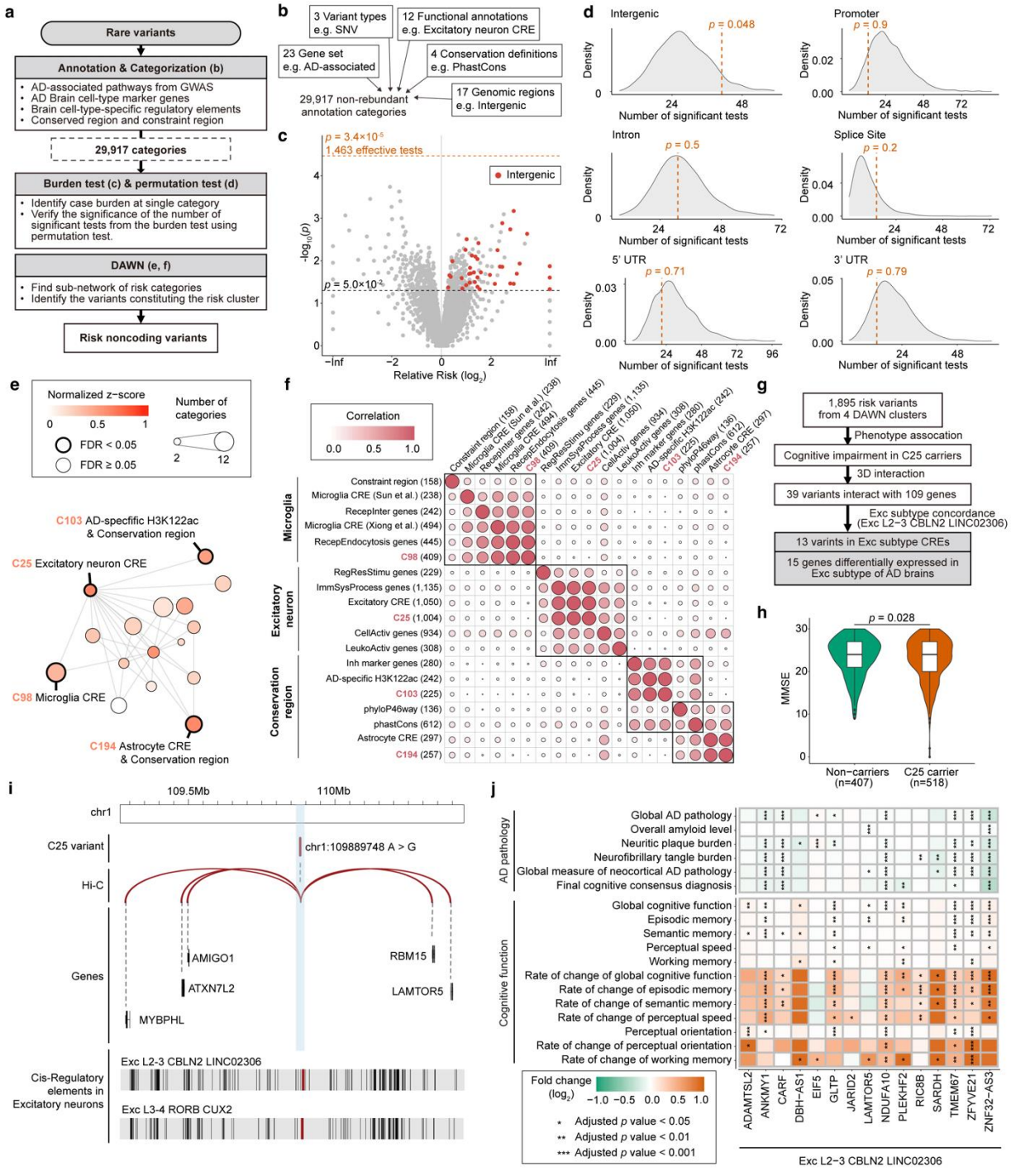
[**Added Discussion, page 28, lines 519–522**]
However, due to the limited statistical power of the CNV analysis, these results such as the *HPSE2* deletion should be cautiously interpreted, and further validation in larger datasets or independent cohorts is necessary to confirm their significance.

[**Added the Abstract, page 3, lines 69–70**]
Moreover, structural variation analysis showed that short tandem repeat expansion was associated with an increased risk of AD.

6. CWAS section. The methodology used in the CWAS section still lacks methodological justification and/or references. It seems like the double-dipping is still a potential issue. Permutation test does not fix the double-dipping issue. The Lasso and the burden test cannot be conducted on the same sample.
Response: We understand the reviewer's concern. To resolve the issue of double-dipping, we excluded the Lasso analysis and only performed the burden test as in our previous studies[1,2] and other recent studies[6]. Although this resulted in slight differences compared with the previous results, the significant signal in the intergenic region was replicated (page 15, lines 260–268).

In the subsequent sub-network analysis (DAWN), we identified four risk clusters (FDR < 0.05 & RR > 1). Among them, cognitive impairment was observed in case carriers of C25, with variants located in the excitatory neuron (Exc) cis-regulatory elements (CREs). To further examine the variants in C25, we investigated their 3D interaction genes and the Exc subtypes in which the variant-associated CREs were located. Specifically, we identified genes that interacted physically with these CREs and assessed whether these genes exhibited differential expression in the corresponding Exc subtypes. Through this approach, we identified 13 variants and 15 genes in the Exc L2–3 CBLN2 LINC02306 subtype, previously identified as an Exc subtype associated with cognitive function[7]. Based on these findings, we revised the description (pages 15–16, lines 269–290) and updated Figures 3, 5 and Extended Figs. 17,18.

[Modified Fig. 3]



[Modified the Fig. 3 b, c, d, f, h, i, j legend]

b, Each variant was annotated with 59 terms across five annotation groups, resulting in 29,917 non-redundant annotation categories.

c, Volcano plot showing enrichment tests for each category using the binomial exact test. Intergenic categories are colored in red (RR > 1, p < 0.05).
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. Permuted expected distribution is shown in gray.
f, The correlation between single annotations and four risk clusters (FDR < 0.05, RR > 1). The four risk clusters were grouped into three terms based on correlation > 0.5. The numbers of variants are shown in parentheses.
h, Comparison of K-MMSE scores between Cluster 25 variant carriers and non-carriers in the Aβ-positive group.
i, Example of prioritized risk variants from Cluster 25. The variant interacted with several genes and was located at specific excitatory subtype regulatory elements.
j, Heatmap depicting the expression of genes associated with risk variants across diverse AD phenotypes. Each color gradient represents the magnitude of log2-fold change, with significance denoted by *. Genes are listed at the bottom. Only gene-variant pairs corresponding to the excitatory neuron subtype, Exc L2–3 CBLN2 LINC02306, are included in the heatmap.

[**Modified Results, pages 15–16, lines 246–247, 260–290**]

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

We compared the burden across each of the 29,917 categories. Although no category showed enrichment after multiple testing corrections (**Supplementary Table 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, Supplementary Table S3c**). Applying the same analysis to clinical diagnosis did not reveal significant signals in the noncoding regions. Power calculations indicated limited power for clinical diagnosis, directing subsequent analyses toward Aβ positivity-associated categories (**Extended 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, Supplementary Table 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[8] (**Supplementary Table 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, Extended 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 patients (p = 0.028, **Fig. 3h**). C25 comprises variants located in excitatory neuron CREs associated with cognitive function in AD in a previous study[7]. We examined the interactions between each variant and gene(s) using Hi-C data to identify the genes affected by these variants[9]. 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[8] (**Fig. 3i**) and examined the association between gene expression and phenotypic characteristics within the same excitatory subtypes using available AD single-cell transcriptome data[7]. 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**, **Supplementary Table S3f**).

[Modified Fig. 5]

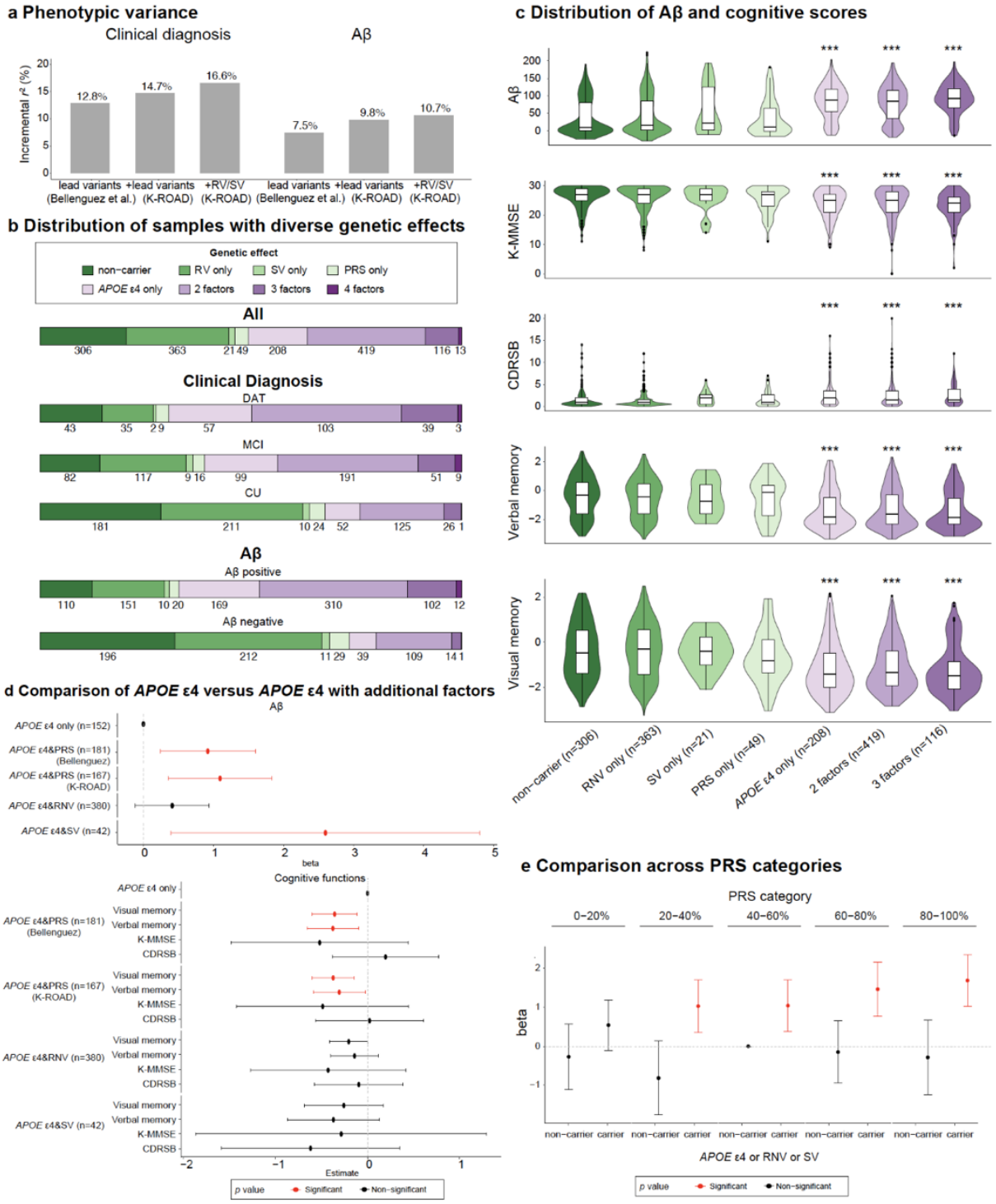
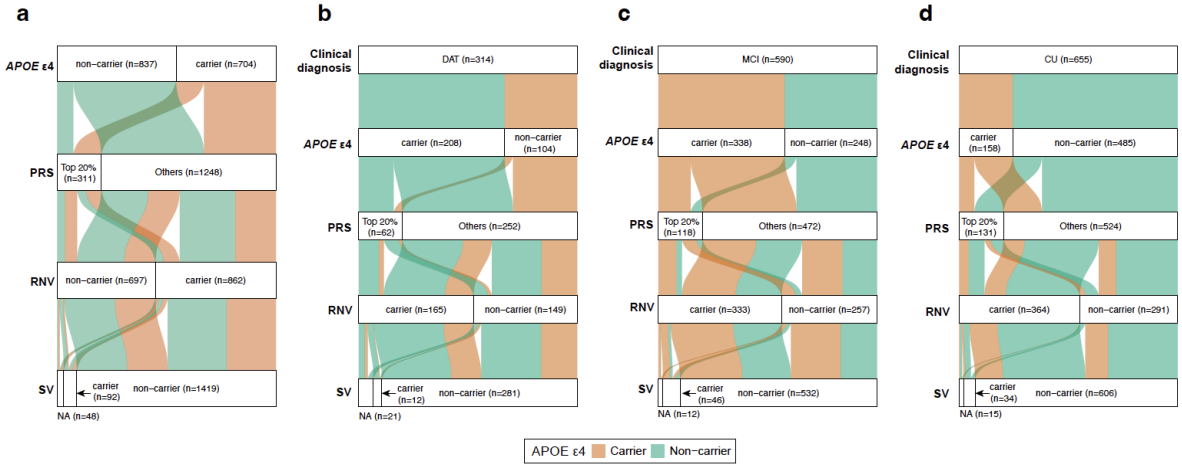


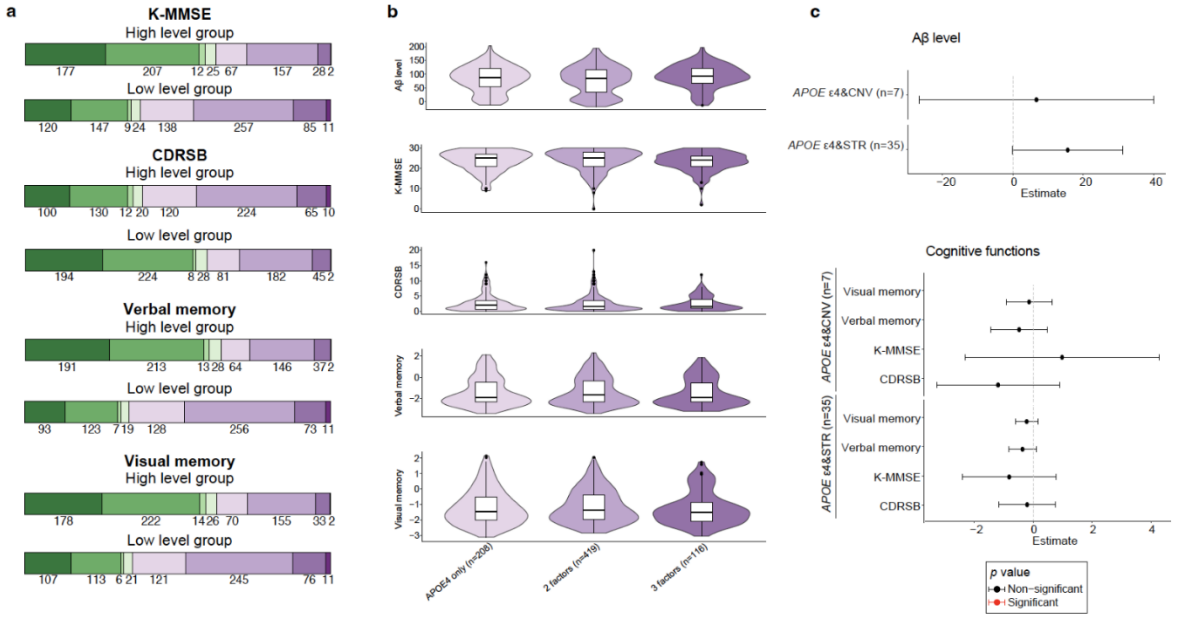
Fig. 5. Diverse effects of genetic variants on Aβ levels and cognitive function.

[Modified to Extended Fig. 17]



Extended Fig. 17. Distribution of APOE ε4 carriers, individuals in the high PRS group, RNV, and SV carriers.

[Modified to Extended Fig. 18]



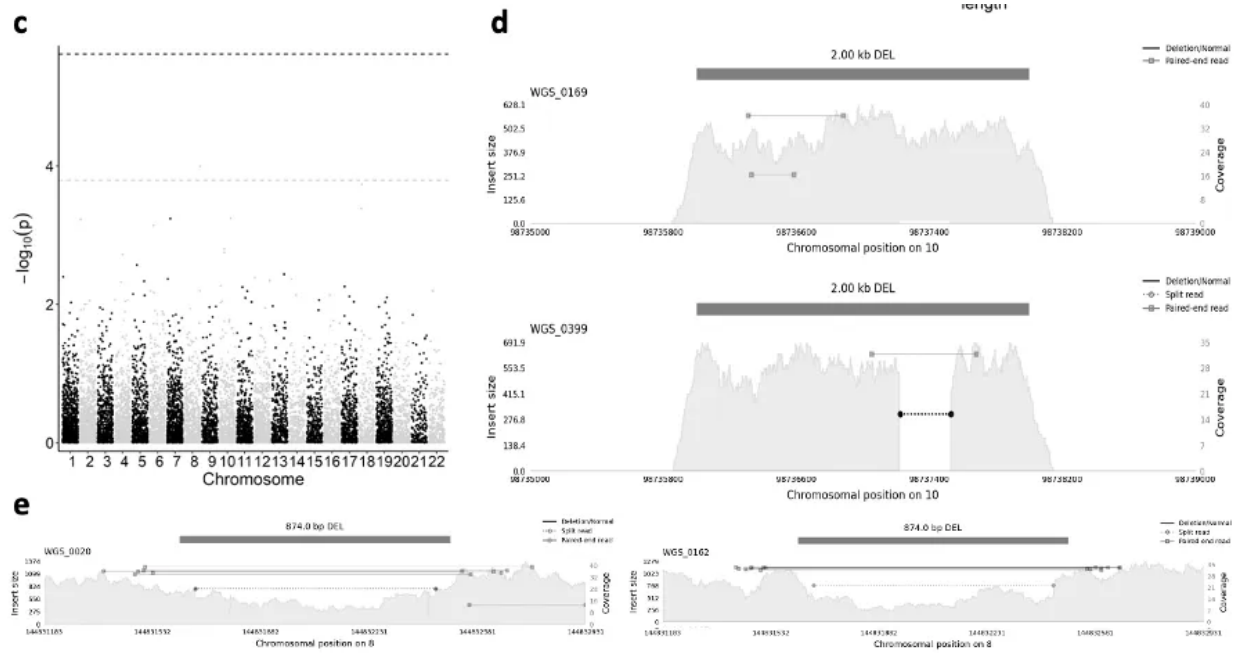
Extended Fig. 18. Distribution of genetic effects on cognitive function and comparison of the variant types.

7. Also, did the authors adjust for covariates in the Lasso regression? The lasso is typically used for predictive models, and post-selection inference with Lasso requires specific statistical procedures (for example, PMID: 25574062).
Response: Thank you for your comments. As described in our response to the comment 6, we excluded the results of the Lasso analysis from this study to avoid the double-dipping issue and ensure methodological robustness.

8. Related to the point above, in the CWAS analysis, 72 categories are first selected for nominal significance, then 13 are identified as belonging to each of intergenic and splice site regions, and these are pass evaluation based on a permutation test. Then, in the next step, all 496 intergenic categories are used for the network analysis. What is the rationale for expanding from the 13 to all 496?
Response: In the CWAS analysis, the permutation test revealed that intergenic categories had a significantly higher number of nominally significant associations; however, none passed the multiple testing threshold. This suggests that the intergenic region as a whole contributes to AD risk, but individual annotation categories do not exhibit strong associations independently. Based on this observation, DAWN was applied to examine how intergenic risk is organized at the network level. Since the DAWN evaluates posterior probability based on enrichment in neighboring clusters within the simulated P-value correlation network, including all intergenic categories (rather than only the nominally significant ones) allows for a more comprehensive network construction that captures potential interactions even among categories that appear insignificant when analyzed individually[1].

9. In response to reviewer #3–point #5, the authors claim they conducted additional analysis investigating the Samplot, but the data is not provided to the reviewer.
Response: We apologize for not including this information in the point-by-point response. Although the samplot data were added to Extended Fig. 12d, 12e, we did not explicitly mention these. For clarity and completeness, we are now providing the data here.

[**Modified to Extended Fig. 12d, 12e**]

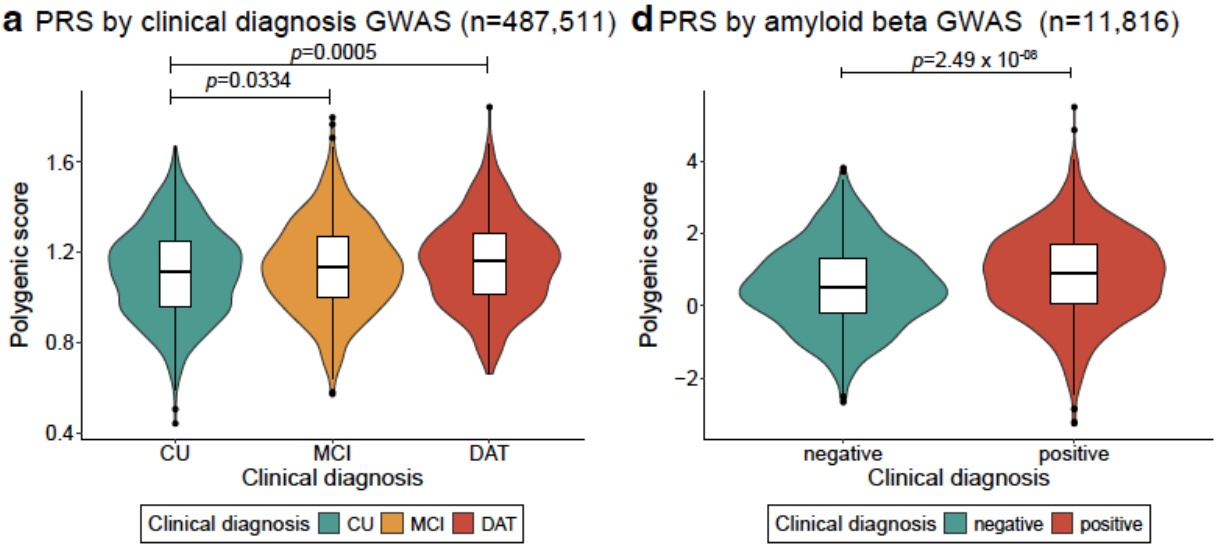


Extended Fig. 12. Characteristics of copy number variants. **d**, Differences between HPSE2 deletion carriers and non-carriers confirmed through samplot. **e**, Ambiguity between ZNF7 false positive deletion carriers and positive deletion carriers confirmed using samplot.

10. The authors add a citation regarding the reviwers concerns about the validity of transferring PRS models between populations. However, the closeness to the threshold of the amyloid-beta PRS (now Extended Figure 15d) is not addressed.

Response: We fully agree with the reviewer's concern regarding the p-value's proximity to the threshold. The GWAS used for the amyloid-beta PRS involved 11,816 individuals, which is relatively smaller than the clinical diagnosis GWAS (n = 487,511), potentially explaining the p-value's closeness to the threshold. However, after adjusting for confounding factors, such as age, sex, batch, education, and the principal components (PC1 to PC10), we observed more significant p-values (p = 2.49 x 10^{-8}). Initially, we presented the p-values obtained from a simple test in the violin plot; however, these have now been updated following the adjustments, and the Results, Methods, and Figure legend have been revised accordingly. Furthermore, the amyloid-beta GWAS-derived PRS demonstrated robust associations with clinical diagnosis ($p = 5.65 \times 10^{-8}$), cognitive functioning based on 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 ($p = 3.36 \times 10^{-5}$), reinforcing the reliability of the findings (page 23, lines 411–421; page 41, lines 886–891).

[**Modified Extended Fig. 16 a,e**]



Extended Fig. 16. Comparison of performance between clinical diagnosis and Aβ GWAS-derived PRS. **a**, Comparison of PRS calculated from clinical diagnosis GWAS (n = 487,511) with clinical diagnosis. **d**, Comparison of PRS calculated from Aβ GWAS (n = 11,816) with Aβ. Significance was calculated by linear regression for AD pathology and logistic regression for cognitive function. Error bars represent the 95% confidence interval.

[**Modified to the Results, page 23, lines 411–421**]

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) (**Extended Fig. 16a**). Clinical diagnosis (p = 0.0005) and Aβ positivity (p = 0.0282) were significantly correlated with PRS (**Extended 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 (**Extended Fig. 16c**). Aβ-negative samples had lower PRSs—calculated from GWAS based on Aβ positivity—than Aβ-positive samples ($p = 2.49 \times 10^{-8}$) (**Extended Fig. 16d**). Individuals with high Aβ PRS were at a 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 (**Extended Fig. 16e, 16f**).

[**Modified to the Methods, page 41, lines 886–891**]
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.

11. The author find the authors have satisfactorily addressed our concerns with the "additive model" of disease by renaming it as "cumulative effects." They additionally perform a new incremental r^2 analysis; I find this analysis informative, and the reviewer appreciate the effort.

Response: We appreciate the reviewer's positive feedback and are grateful for the recognition of our efforts in renaming the "additive model" to "cumulative effects" and performing the new incremental r^2 analysis. We are glad to hear your remark that the analysis was informative.

[Minor points (#12 and #14)]

12. It's ambiguous what line 117 is saying about the p-values. Is this 2.2e-17 or 2.2 raised to the negative 17th power?

Response: We apologize for the confusion. The value referred to is 2.2e-17, and for clarity, we revised it to ($p < 2.2 \times 10^{-16}$).

[**Modified to the Results, page 6, lines 115–117**]
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.

13. Line 140 has two sentences spliced together.

Response: We apologize for the error in sentence structure. The sentence has been revised and split into two for clarity.

[**Modified to the Results, page 7, lines 138–141**]
The Japanese cohort comprised 140 individuals 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.

14. Lines 141-142: Please clarify which dataset, Korean WGS or the chip data, the description “523 DAT and 340 CU individuals” refers to.

Response: We apologize for the confusion. We intended to emphasize that the samples were independent from the Korean WGS, but the phrasing may be unclear. We revised the sentence to improve clarity.

[**Modified to the Results, page 7, lines 139–141**]
The Korean cohort included 523 DAT individuals and 340 CU individuals, all from independent samples, and genotyping was conducted using a chip array.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: We appreciate the reviewers' valuable time and effort in reviewing our manuscript.

Reviewer #5 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: We appreciate the reviewers' valuable time and effort in reviewing our manuscript.

Response letter references

1. An, J.Y. *et al.* Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**(2018).
2. 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).
3. 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).
4. Prokopenko, D. *et al.* Whole-genome sequencing reveals new Alzheimer's disease-associated rare variants in loci related to synaptic function and neuronal development. *Alzheimers Dement* **17**, 1509-1527 (2021).
5. Shigemizu, D. *et al.* Whole-genome sequencing reveals novel ethnicity-specific rare variants associated with Alzheimer's disease. *Mol Psychiatry* **27**, 2554-2562 (2022).
6. Gillani, R. *et al.* Rare germline structural variants increase risk for pediatric solid tumors. *Science* **387**, eadq0071 (2025).
7. 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 e27 (2023).
8. Xiong, X. *et al.* Epigenomic dissection of Alzheimer's disease pinpoints causal variants and reveals epigenome erosion. *Cell* **186**, 4422-4437 e21 (2023).
9. 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).

Manuscript Number: NCOMMS-24-30438B

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

Responses from the Authors for Review Comments:

We appreciate the reviewers' thoughtful and constructive feedback. In this study, we comprehensively analyzed various genetic variations using whole-genome sequencing data from Korean individuals with and without Alzheimer's disease. Our findings include associations between genetic variations and AD-related endophenotypes, such as amyloid positivity and cognitive testing scores. Additionally, we replicated findings from a recent European study that investigated short tandem repeats in patients with AD.

In response to the reviewers' comments, we revised and refined our description in the manuscript. We believe these revisions have enhanced the quality and clarity of the study.

REVIEWER COMMENTS

Reviewer #3 (Remarks to the Author):

The authors have made substantial improvements in the revised manuscript. The reviewer appreciates the clear formatting of both the response letter and the revised document, which made it very easy to review the updated manuscript.

The reviewer has two minor comments:

1. Abstract. The statement "Moreover, structural variation analysis showed that short tandem repeat expansion was associated with an increased risk of AD" should acknowledge that the CNV association at the HPSE2 locus has borderline statistical significance at the best rather than being strongly associated.

Response: We appreciate the reviewer's insightful comment. In accordance with the suggestion, we have revised the Abstract to temper the interpretation of the structural variation finding. Specifically, we have modified the relevant sentence to clarify that the CNV at the HPSE2 locus shows only borderline statistical significance.

[Modified to the Abstract]

Moreover, structural variant analysis revealed that short tandem repeat expansions were significantly associated with an increased risk of AD, and copy number variant at the HPSE2 locus showed borderline statistical significance.

2. Data Availability Section. The reviewer encourages the authors to adhere to open data principles and ensure that resources are readily accessible to the research community. For example, instead of stating that "Korean scRNA datasets are available from the corresponding author," the data should be deposited in a standard repository commonly used in the field. Similarly, the GWAS catalog accession ID for the new GWAS summary statistics should be included in the published version of the manuscript.

Response: We appreciate the reviewer for the suggestion to follow open data principles. In response, we have deposited the GWAS summary statistics generated in this study to the NHGRI-EBI GWAS Catalog under accession number at GCST90566388 and

GCST90566389 (https://www.ebi.ac.uk/gwas), and we have added this information to the Data Availability section of the manuscript.

For the Korean single-nucleus RNA-seq data, due to the sensitivity of the data and the need to protect participant privacy, we are unfortunately unable to make the raw data publicly available. However, to promote transparency and enable collaborative research, we have included detailed instructions in the Data Availability section regarding how researchers can request access. Specifically, data access requests may be directed to the corresponding author (S.W. Seo at sangwonseo@empas.com), and all requests will be reviewed and responded to within 2 to 4 weeks.

[Modified to the Data availability]

The GWAS summary statistics for Koreans generated in this study have been deposited in the NHGRI-EBI GWAS Catalog under accession number at GCST90566388 and GCST90566389 [https://www.ebi.ac.uk/gwas]. ...(omited) The single-nuclei RNA-seq data for Koreans are accessible for collaborative research under restricted conditions to protect participant privacy. For inquiries regarding data access, please contact the corresponding authors (S.W. Seo (sangwonseo@empas.com)). Data access requests will be reviewed and responded to within 2 to 4 weeks of receipt.

These are minor changes, and the reviewer believes the editors can determine whether the authors have sufficiently addressed them—there is no need for further reviewer feedback. Congratulations on bringing such a complex and valuable project to completion.

Response: We sincerely thank the reviewer for the thoughtful comments and kind encouragement. We are grateful that the constructive feedback provided throughout the review process helped us further improve the clarity and quality of our work, and we deeply appreciate all the reviews that contributed to strengthening the manuscript.