

Genome-wide analysis of common structural variants in Alzheimer's disease identifies a risk-associated regulatory CCG repeat in *FAM193B*

Received: 6 May 2025

Accepted: 3 September 2026

Published online: 12 October 2026

 Check for updates

Han Cao ^{1,2}, Yifan Fu ¹, Xiaopu Zhou ^{1,2,3}, Yangyang Duan ^{1,2}, Fanny C. F. Ip ^{1,2,3}, Yu Chen^{1,3,4,5}, Shenyuan Hu ⁶, Si Chai Liu ², Timothy C. Y. Kwok ⁷, Kin Y. Mok^{1,2,8}, Tom H. Cheung ^{1,2,3,6}, John Hardy^{2,8,9}, Amy K. Y. Fu ^{1,2,3,4} ✉ & Nancy Y. Ip ^{1,2,3,4} ✉

Although Alzheimer's disease (AD) has a polygenic architecture, with numerous associated single-nucleotide polymorphisms (SNPs) identified, the contribution of structural variants (SVs) is unclear. In this study, using long-read ($n = 11$) and short-read ($n = 3,190$) sequencing data from the Chinese population, we show that common SVs contribute to AD heritability beyond common SNPs. We identify an expanded CCG repeat in the *FAM193B* promoter that shows a genome-wide significant association with increased AD risk (odds ratio = 2.20 per 10-repeat increase, $P = 5.5 \times 10^{-13}$). The expanded *FAM193B*-CCG repeat increases *FAM193B* expression and mediates the pleiotropic association between brain *FAM193B* expression and AD risk. Analyses of postmortem human brains and cultured neuronal cells show that elevated neuronal *FAM193B* expression is associated with DNA damage and cell death-related pathways. These findings establish the common *FAM193B*-CCG repeat as an AD risk genetic factor and highlight the contribution of common SVs to AD heritability.

AD, a progressive neurodegenerative disorder, is the most common type of dementia and one of the leading causes of death in the aged population¹. AD is highly heritable and exhibits polygenic architecture². Genome-wide association studies (GWASs) have identified more than 70 risk loci for AD³. Most identified AD risk genetic variants are common variants with minor allele frequency (MAF) > 0.05, such as *APOE ε4*, which is the most significant AD risk genetic variant and has high disease penetrance (odds ratio (OR) = 3.2 for heterozygous carriers)⁴. Although other common genetic variants exert relatively weak effects on AD risk, their polygenic effects account for a substantial proportion of heritability⁵. Nonetheless, the known AD-associated genetic variants do not fully explain the observed heritability of AD², indicating the presence of unidentified disease-associated genetic factors.

Existing genetic studies are limited by their focus on small genetic variants, such as SNPs. Consequently, the disease associations of SVs, which affect genomic regions larger than 50 base pairs, are largely overlooked⁶. Compared to SNPs, SVs affect more nucleotide content in the human genome and play a role in disease susceptibility⁷. Indeed, genetic studies illustrate strong risk effects of rare SVs on AD and other neurodegenerative diseases^{8–11}. For example, duplication of *APP* is reported to cause autosomal dominant AD in a few families¹¹. Recent genetic studies on SVs and tandem repeat expansions reveal an increased polygenic burden of rare SVs in patients with sporadic AD^{12,13}. These findings indicate that rare SVs modulate the risk of developing AD. However, it is largely unclear how common SVs contribute to AD risk. Given the high prevalence of common SVs, investigating their roles

A full list of affiliations appears at the end of the paper. ✉ e-mail: boamy@ust.hk; boip@ust.hk

in modulating AD risk and gene expression is important to understand the genetic etiology of AD in the general population. Moreover, as SVs generally exert stronger effects on molecular phenotypes than individual SNPs¹⁴, they present new opportunities for identifying common genetic variants with relatively high AD penetrance.

The major bottleneck to studying SVs has been reliance on microarray or short-read sequencing (SRS) technologies, which cannot accurately discover or genotype more than half of SVs owing to their limited ability to resolve full sequences spanning the SVs¹⁵. However, recent advances in long-read sequencing (LRS) enable sensitive discovery of SVs using continuous long reads of several kilobase pairs^{16–18}. Furthermore, LRS provides precise information about SV sequences and breakpoints, enabling the use of SRS data to genotype SVs with relatively high accuracy¹⁹.

Accordingly, in the present study, we used a hybrid sequencing strategy combining small-scale LRS data and large-scale SRS data to investigate the contribution of common SVs to AD, thereby mitigating the high cost of LRS. We show that common SVs contribute to additional AD risk beyond that explained by common SNPs. Notably, we identified a common CCG repeat polymorphism in the *FAM193B* promoter as an AD risk variant with genome-wide significance. This *FAM193B*-CCG repeat mediates the pleiotropic association between brain *FAM193B* expression and AD risk. Supporting its regulatory effect, longer *FAM193B*-CCG repeats enhance transcription factor binding to the *FAM193B* promoter and upregulate *FAM193B* gene expression. Moreover, the repeat is associated with neuronal death-related pathways and decreased gray matter volume in the human brain. Consistently, *FAM193B* overexpression in cultured neuronal cells increases DNA damage pathways, suggesting that elevated *FAM193B* expression promotes neuronal death. Thus, our study identifies a common repeat polymorphism in the *FAM193B* promoter as a genetic risk factor for AD and highlights the impact of common SVs on neurodegeneration-related pathways.

Results

SV calling and genotyping

We first generated a reference set of SVs based on the PacBio continuous long-read data from 11 Chinese genomes, including two samples sequenced in-house and nine public samples (Supplementary Table 1). After filtering out low-quality and duplicate reads, we obtained 1.37 terabases of clean long reads with an average length of 10.35 kilobases, which is 40-fold (s.d. = 17.7) average coverage of the human genome (Supplementary Table 2 and Supplementary Fig. 1).

To perform sensitive SV discovery using LRS data, we performed SV calling against the GRCh38 reference genome for each sample by combining two long-read aligners—NGMLR¹⁶ and minimap2 (ref. 20)—with three SV callers: Sniffles2 (ref. 16), cuteSV¹⁸ and SVIM¹⁷ (Fig. 1a). After merging the results from the six SV calling methods, we identified 125,530 non-redundant SVs (Supplementary Table 3). To obtain a high-confidence SV call set, we utilized a random forest model to classify individual SV calls (Fig. 1a), yielding areas under the receiver operating characteristic curves (AUROCs) of 0.981 for insertions and 0.958 for deletions (Extended Data Fig. 1a). Moreover, compared to single SV calling methods, the random forest model enabled more accurate SV calling with scalable stringency (Fig. 1b,c and Supplementary Table 4). By retaining SVs with a false discovery rate (FDR) ≤ 5%, we identified 44,122 SVs, including 21,821 insertions and 22,301 deletions (Extended Data Fig. 2). We compared the filtered SVs with those reported in other studies or represented in the T2T-CHM13 reference genome (Supplementary Table 5). Most common SVs (96.7%) were present in public SV datasets, whereas the novel SVs identified in this study are predominantly restricted to 1–2 haplotypes (Extended Data Fig. 2b,c), supporting the robustness of the identified SVs. Further demonstrating the accuracy of SV genotypes, in the two parent–child trios, the filtered SVs exhibited lower Mendelian error

rates than raw SV calls (decreasing from 7.7% to 2.4% and from 6.3% to 2.1% in these families; Fig. 1d). Thus, these results demonstrate that our multi-method SV calling approach yields high-confidence SVs.

We subsequently genotyped the high-confidence SVs in a Hong Kong Chinese cohort ($n = 2,333$; Fig. 1a and Supplementary Table 6) using Paragraph with high-coverage SRS data (35×)¹⁹. To evaluate the accuracy of SV genotyping, we compared the correlations between the allele frequencies derived from SRS and LRS data. After conventional genetic quality control (Methods), the allele frequencies obtained from SRS data were moderately correlated with those from LRS data for both insertions ($r = 0.693$, $P < 2.2 \times 10^{-16}$; Fig. 1e) and deletions ($r = 0.583$, $P < 2.2 \times 10^{-16}$; Fig. 1f), suggesting the presence of genotyping errors in SRS data. To address these genotyping errors, we performed SRS-based genotyping in five LRS samples that also had 35× SRS data and applied a stringent filtering procedure by retaining SVs that exhibited 100% concordance between the LRS and SRS genotypes in those five samples. The remaining SVs ($n = 17,294$) exhibited strong allele frequency correlation between LRS and SRS data for both insertions ($r = 0.930$, $P < 2.2 \times 10^{-16}$; Fig. 1e) and deletions ($r = 0.942$, $P < 2.2 \times 10^{-16}$; Fig. 1f). We then benchmarked the filtered SVs by performing PacBio HiFi sequencing (30×) on one random sample, using the consensus call from Sniffles and cuteSV across the 17,294 SVs as the truth set. The filtered SRS-based SV genotyping achieved high accuracy, with F1 scores of 0.938 for insertions and 0.934 for deletions (Extended Data Fig. 1b). Therefore, by applying a hybrid sequencing method, we identified and genotyped high-confidence SVs in the study cohort.

Functional annotation of SVs

To investigate the biological functions of the high-confidence SVs identified by LRS data ($n = 44,122$), we examined their enrichment in functional genomic elements. Both insertions and deletions were rarer in exons, 3' untranslated regions (UTRs), regions with distal enhancer-like signatures and CTCF-binding regions ($P < 0.001$; Fig. 2a and Supplementary Table 7). Moreover, SV enrichment analysis in different types of genes revealed that deletions were less prevalent in protein-coding genes (OR = 0.85, 95% confidence interval (CI): 0.76–0.93, $P < 0.001$) and more prevalent in pseudogenes (OR = 3.77, 95% CI: 2.65–4.92, $P < 0.001$) (Fig. 2b and Supplementary Table 8). Therefore, the relative intolerance of SVs in functional regions suggests that they modulate gene expression and functions, potentially conferring risk effects on human diseases.

We subsequently investigated the potential involvement of SVs in human diseases by examining their abundance in known disease risk loci²¹. Notably, neurological disorder loci harbored more SVs ($n = 3,144$, averaging 2.9 SVs per locus) than other disease loci (Fig. 2c). This difference is possibly attributable to the longer genomic spans of neurological disorder loci (Fig. 2d), which reflects the generally greater length of genes involved in neuronal functions²². However, after normalization by locus length, cancer risk loci exhibited the highest SV density (16.1 SVs per megabase pair; Supplementary Table 9). Next, we analyzed the pathogenic impacts of SVs based on their genomic features by using Combined Annotation Dependent Depletion scores for SVs (CADD-SV)²³. SVs with higher CADD-SV deleteriousness scores were more likely to reside in known disease risk loci (Fig. 2e). Notably, among SVs with a CADD-SV score greater than 16 (that is, the top 2.5% most deleterious SVs; $n = 111$), 61.3% of SVs, including 13.2% predicted as loss-of-function SVs, were not in any known risk locus. Thus, these SVs could tag disease risk loci not investigated in previous GWASs.

Polygenic effects of common SVs on AD

Although a few rare SVs are associated with AD^{11–13}, it is unclear how common SVs (that is, MAF > 5%) contribute to AD risk. Therefore, we performed genome-wide association analysis (Supplementary Fig. 2) and polygenic risk score (PRS) analysis using common SNPs or SVs. The PRS models were generated by the clumping and thresholding

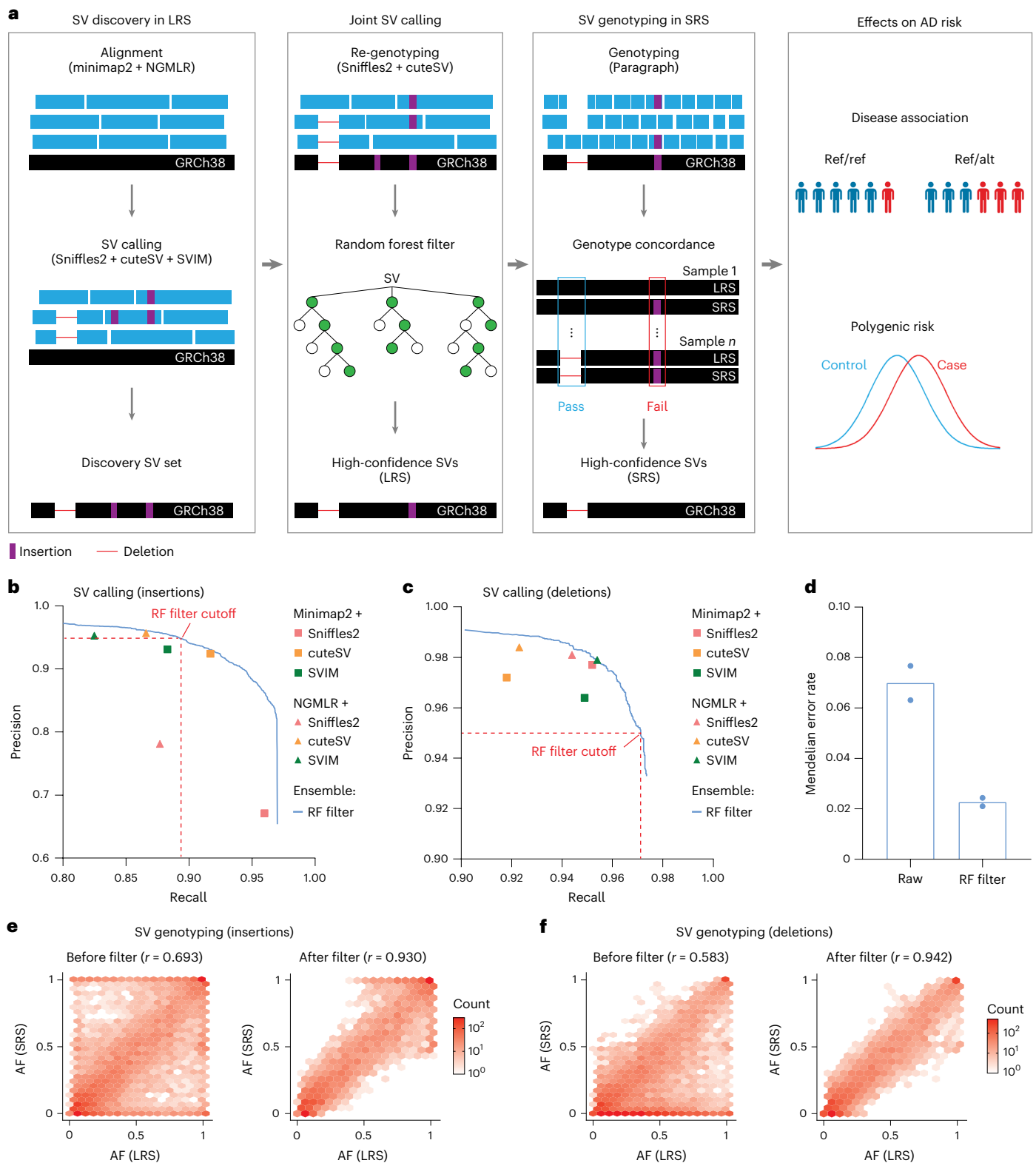


Fig. 1 | SV calling and genotyping using a hybrid sequencing strategy. **a**, Analytical workflow of SV discovery, joint SV calling, SV genotyping and AD association analysis of SVs. Blue and black rectangles denote the sequencing reads and reference genome, respectively. **b**, **c**, Precision–recall curves of insertions (**b**) and deletions (**c**) identified by different SV calling methods.

The random forest filter cutoff was determined by precision = 0.95. **d**, Mendelian error rates of identified SVs. $n = 2$ families. Bars represent mean values. **e**, **f**, Correlations between the allele frequency (AF) obtained from the LRS data and SRS data for insertions (**e**) and deletions (**f**). Alt, alternative allele; r , Pearson's correlation coefficient; Ref, reference allele; RF, random forest.

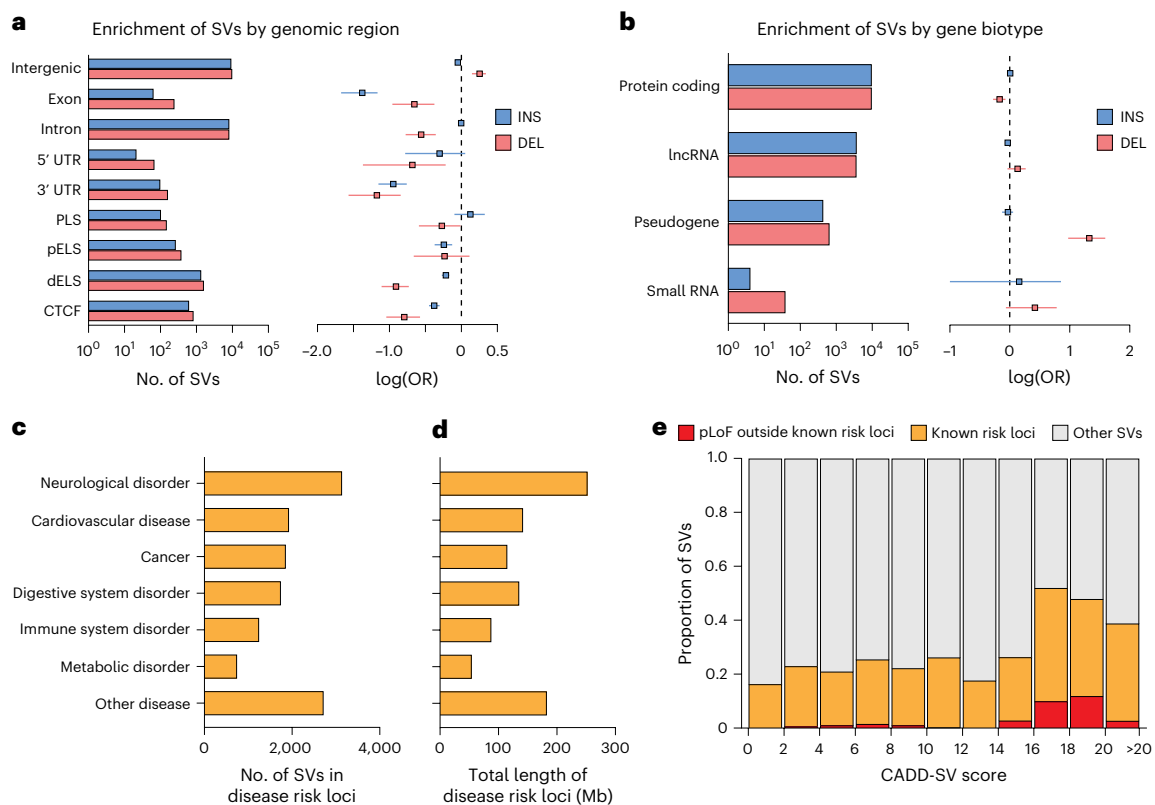


Fig. 2 | Functional annotations of SVs. a, b, Enrichment analysis of SVs by genomic region (**a**) and gene biotype (**b**). Bootstrap method with 1,000 replicates. Data are presented as log(OR) with 95% CI. **c**, Numbers of SVs residing in known disease risk loci. Disease risk loci were defined as the 10-kilobase flanking regions of disease risk genes reported in the GWAS catalog. **d**, Total

length of known disease risk loci. **e**, Predicted deleteriousness of SVs by CADD-SV. DEL, deletion; dELS, distal enhancer-like signature; INS, insertion; lncRNA, long non-coding RNA; Mb, megabase; pELS, proximal enhancer-like signature; pLoF, predicted loss of function; PLS, promoter-like signature.

method with repeated 10-fold cross-validation, excluding the *APOE* locus to prevent it from dominating the models. The PRS models of SNPs and SVs best predicted AD with *P* value thresholds of 0.01 and 0.05, respectively (Supplementary Table 10). Interestingly, the SV models predicted AD risk ($r^2 = 0.056$, AUROC = 0.611) better than the SNP models ($r^2 = 0.046$, AUROC = 0.597) ($P = 0.022$) (Fig. 3a,b). Moreover, the PRSs calculated from SNPs and SVs were weakly correlated ($r^2 = 0.005$; Fig. 3c), suggesting that SNPs and SVs capture different genetic risks of AD. Accordingly, we constructed PRS models combining both SNPs and SVs ('SNP + SV models' hereafter), which explained more AD variance ($r^2 = 0.098$, AUROC = 0.649) than the models solely using SNPs or SVs ($P \leq 2.6 \times 10^{-15}$) (Fig. 3a,b). To investigate whether the improved performance of the SNP + SV models was due to the linkage disequilibrium between SVs and SNPs, we conducted conditional PRS analysis ('SNP_{adj} + SV models' hereafter) by adjusting the SNPs' effect with the lead AD-associated SV within 1 megabase. The SNP + SV and SNP_{adj} + SV models showed similar predictive performance for AD ($P > 0.11$; Extended Data Fig. 3a,b) and yielded highly correlated PRSs ($r^2 = 0.977$, $P < 2.2 \times 10^{-308}$; Extended Data Fig. 3c), suggesting that the SNP + SV models' performance is not confounded by linkage disequilibrium between SNPs and SVs.

To validate the PRS models, we applied a representative SNP + SV model (that is, one with a PRS r^2 value close to the mean of the cross-validation results) to AD and cognitively normal controls (NCs) in the corresponding validation set as well as additional participants with mild cognitive impairment (MCI). Accordingly, individuals with MCI had higher PRSs than NCs but lower PRSs than patients with AD (Fig. 3d). We then categorized participants into low, medium and high PRS groups according to PRS tertiles. Interestingly, the high PRS group had larger proportions of individuals with AD (16%) and MCI (75%) than

the medium PRS group (AD: 11%, MCI: 70%) or the low PRS group (AD: 9%, MCI: 62%) (Fig. 3e). Moreover, both the SV model (AUROC = 0.64, 95% CI: 0.58–0.70) and the SNP + SV model (AUROC = 0.62, 95% CI: 0.56–0.68) distinguish MCI from NC more accurately than the PRS model solely based on SNPs (AUROC = 0.54, 95% CI: 0.48–0.60, DeLong test $P \leq 0.012$, net reclassification improvement ≥ 0.146 ; Fig. 3f and Supplementary Table 11), underscoring the significant contributions of common SVs to AD and related dementias. Therefore, the PRS analyses collectively suggest that common SVs account for additional heritability of AD that is not captured by common SNPs.

Identification of a common AD risk repeat polymorphism in the *FAM193B* locus

The significant contribution of common SVs to AD suggests that they are important genetic risk factors. To identify candidate AD-associated SVs that may exert causal effects on molecular functions during disease pathogenesis, we analyzed the associations between AD and common SVs predicted to be highly deleterious (that is, CADD-SV score > 16). Among the deleterious SVs, an insertion in the promoter/5' UTR of the *FAM193B* gene was most significantly associated with AD (OR = 1.60, 95% CI: 1.17–2.21, $P = 3.6 \times 10^{-3}$, CADD-SV score = 19.6; Supplementary Table 12). For validation of this insertion, we performed PacBio HiFi sequencing on a random heterozygous carrier of this insertion. The inserted sequence was identified as an expanded CCG repeat (Fig. 4a), highlighting that the prior analysis, which treated the repeat as a fixed-length insertion, does not fully capture its polymorphism. To quantify repeat polymorphisms, we reanalyzed the SRS data using tandem repeat callers (Methods and Supplementary Fig. 3a)^{24–26}. The *FAM193B*-CCG repeat was highly polymorphic, with alleles ranging from four to more than 30 repeats (Fig. 4b). Moreover, the *FAM193B*-CCG

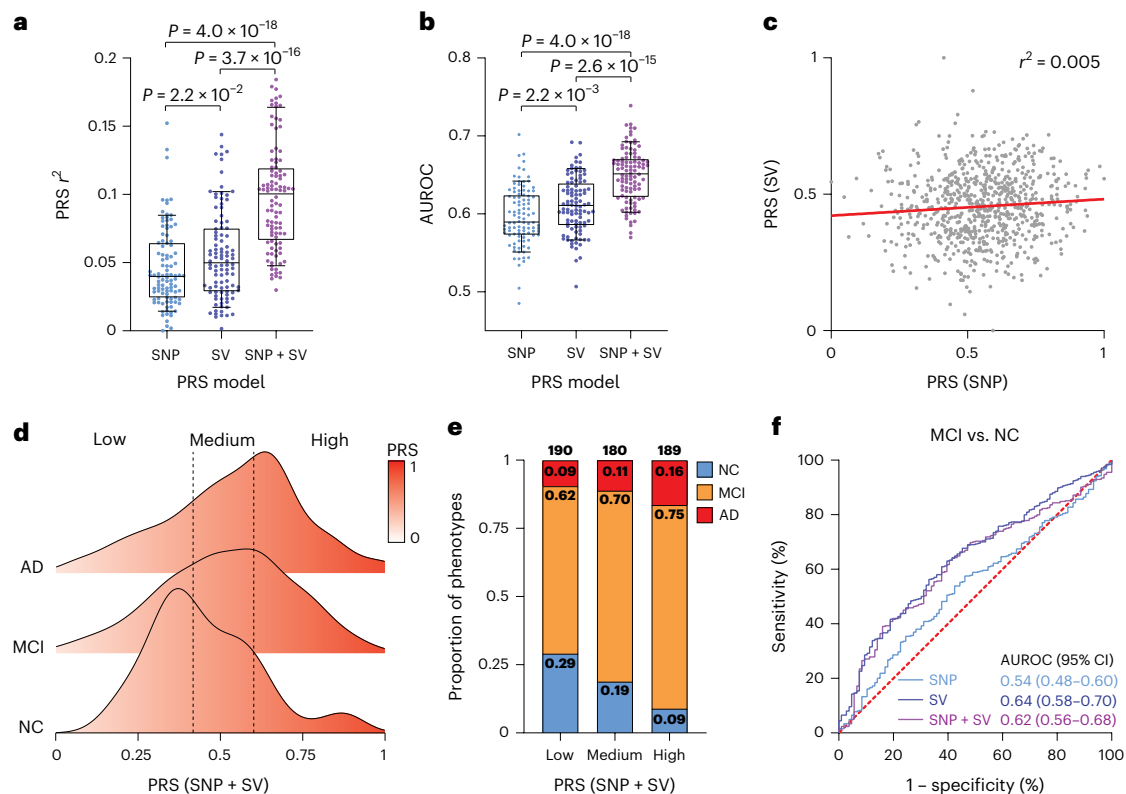


Fig. 3 | Polygenic effects of common SVs on AD. a, Association between AD and PRSs calculated using SNPs, SVs and both SNPs and SVs ('SNP + SV'). Ten times repeated 10-fold cross-validation. Paired two-tailed Wilcoxon signed-rank test. r^2 , Nagelkerke's pseudo r^2 . **b**, AUROCs showing the performance for distinguishing patients with AD from NCs using SNP, SV and SNP + SV PRS models. Ten times repeated 10-fold cross validation. Paired two-tailed Wilcoxon signed-rank test. In **a** and **b**, box plots show medians, interquartile ranges, 10th and 90th percentiles and all individual values. **c**, Correlations between the PRSs calculated from SNPs

and SVs. r^2 , Pearson's correlation coefficient. **d**, Distributions of the PRSs in participants with AD, participants with MCI and NCs. Dashed lines denote PRS tertiles of participants with a low, medium or high risk of AD. **e**, Proportions of NC, MCI and AD participants in different AD risk groups. **f**, ROC curve showing the performance of the PRS model for distinguishing participants with MCI from NCs. The AUROCs and 95% CIs are shown in the bottom right. The PRS models from a representative fold ($r^2 = 0.047, 0.055$ and 0.088 for the SNP, SV and SNP + SV PRS models, respectively) were used in **c–f**.

repeat polymorphisms genotyped by tandem repeat callers were correlated with the high-confidence insertion genotypes generated by the Paragraph method (Supplementary Fig. 3b), indicating that the repeat genotyping results are also of high confidence.

We analyzed the dosage effects of the *FAM193B*-CCG repeat on AD risk by performing association analysis using both the additive model (that is, the summed repeat length from both alleles) and the dominant model (that is, the repeat length of the longer allele). Accordingly, the *FAM193B*-CCG repeat was significantly associated with increased AD risk in both the additive model ($\beta = 0.056, P = 5.3 \times 10^{-8}$) and the dominant model ($\beta = 0.081, P = 8.1 \times 10^{-10}$) ($n = 651$ AD versus $n = 981$ NC; Fig. 4c and Supplementary Table 13). Conditional analysis suggests that the effect of the *FAM193B*-CCG repeat is independent of nearby SNPs ($P = 4.3 \times 10^{-7}$ and $P = 6.9 \times 10^{-9}$ for the additive and dominant models, respectively; Supplementary Table 14). Furthermore, association analysis using an independent mainland Chinese cohort (Supplementary Table 6)²⁷ replicated the association between the *FAM193B*-CCG repeat and increased AD risk (additive model: $\beta = 0.035, P = 7.7 \times 10^{-3}$; dominant model: $\beta = 0.073, P = 3.0 \times 10^{-4}$; $n = 865$ AD versus $n = 693$ NC; Supplementary Table 13). Meta-analysis of the two Chinese cohorts identified genome-wide significant associations (Bonferroni correction cutoff = 1.7×10^{-8}) between the repeat and AD for both the additive model ($\beta = 0.048, P = 1.2 \times 10^{-9}$) and the dominant model ($\beta = 0.079, P = 5.5 \times 10^{-13}$) ($n = 1,516$ AD versus $n = 1,674$ NC; Supplementary Table 13).

Of note, the estimated per-repeat effect sizes of the *FAM193B*-CCG repeat suggest that long repeat expansions may confer relatively high

disease penetrance (equivalent to ORs of 1.6 and 2.2 per 10 repeats for the additive and dominant models, respectively). To estimate the risk effects of *FAM193B*-CCG with different repeat length cutoffs, we performed association analysis of AD by comparing participants with the longer allele carrying $\geq n$ CCG repeats against those who harbored ≤ 11 CCG repeats. Accordingly, the AD risk effects of the *FAM193B*-CCG repeat increased with repeat length (Fig. 4d–f). In particular, the *FAM193B*-CCG repeat with ≥ 18 CCG repeats exhibited the most significant association with AD (OR = 2.11, 95% CI: 1.68–2.65, $P = 1.1 \times 10^{-10}$; Fig. 4f and Supplementary Table 15) and was prevalent in the Chinese population (carrier frequency = 0.199 and 0.309 in NCs and patients with AD, respectively). These findings collectively suggest that the common *FAM193B*-CCG repeat polymorphism is an AD risk variant with genome-wide significance in the Chinese population.

We subsequently replicated the above association analysis in the European population using the UK Biobank (Supplementary Methods)²⁸. Accordingly, the *FAM193B*-CCG repeat exerted nominally significant association with increased AD risk in the European population (additive model: $\beta = 0.015, P = 2.9 \times 10^{-3}$; dominant model: $\beta = 0.012, P = 9.2 \times 10^{-3}$; $n = 2,674$ clinically diagnosed AD versus $n = 22,996$ NC; Supplementary Table 13). Among all common *FAM193B*-CCG repeats, alleles with ≥ 19 CCG repeats (*FAM193B*-CCG₁₉) were most significantly associated with AD (OR = 1.21, 95% CI: 1.05–1.41, $P = 1.1 \times 10^{-2}$; Fig. 4g and Supplementary Table 16), which is similar to the ≥ 18 CCG cutoff in the Chinese population. Conditional analysis showed that the effect of the *FAM193B*-CCG repeat is independent of the lead AD-associated SNP (that is, rs335480) reported in European AD GWAS^{3,29} ($P = 1.6 \times 10^{-2}$

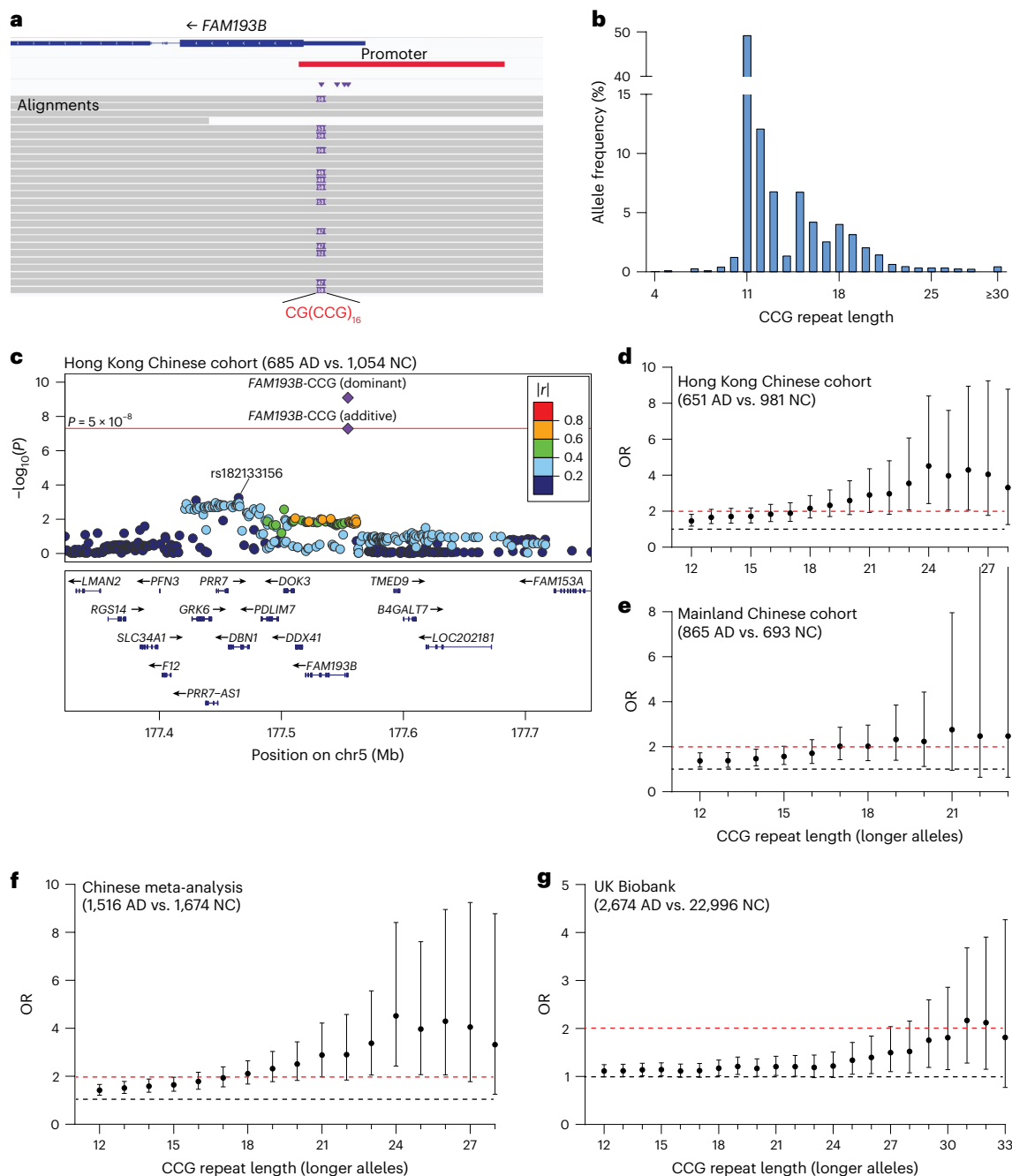


Fig. 4 | Identification of a common AD-associated CCG repeat in the *FAM193B* locus. **a, PacBio HiFi sequencing validation of the insertion in the *FAM193B* promoter region. The promoter annotation (red) was retrieved from the SCREEN database. **b**, Prevalence of the *FAM193B*-CCG repeat polymorphism in the study cohort. **c**, LocusZoom plot showing the associations of genetic variants in the *FAM193B* locus with AD. Logistic regression. Filled colors denote the linkage disequilibrium (as measured by $|r|$) between the *FAM193B*-CCG repeat and nearby genetic variants. **d,e**, Risk effects of the *FAM193B*-CCG repeat on AD with different**

repeat length cutoffs in the Hong Kong Chinese cohort (**d**) and the mainland Chinese cohort (**e**). **f**, Meta-analysis of the risk effects of the *FAM193B*-CCG repeat on AD in the Chinese population. Fixed-effect meta-analysis model. **g**, Risk effects of the *FAM193B*-CCG repeat on AD with different repeat length cutoffs in the UK Biobank. In **d–g**, associations were examined by logistic regression under a dominant model; data are presented as ORs with 95% CI. Black and red dotted lines indicate ORs of 1 and 2, respectively. Mb, megabase.

and $P = 4.5 \times 10^{-2}$ for conditional analysis with additive and dominant models, respectively; Supplementary Table 14). To explore the correlation between the *FAM193B*-CCG repeat and nearby AD-associated SNPs, we performed linkage disequilibrium analysis of *FAM193B*-CCG₁₉₊ with other SNPs in the *FAM193B* locus. Although most SNPs in this locus were not in linkage disequilibrium with *FAM193B*-CCG₁₉₊ ($|r| < 0.2$), the candidate AD-associated SNPs that passed the locus-wise significance threshold ($P < 0.05 / 831 = 6.0 \times 10^{-5}$) exhibited the highest

linkage disequilibrium with *FAM193B*-CCG₁₉₊ ($0.2 < |r| < 0.5$; Extended Data Fig. 4a). Moreover, the AD risk effects of SNPs in the *FAM193B* locus were positively correlated with their linkage disequilibrium (as measured by r) with *FAM193B*-CCG₁₉₊ ($r = 0.550$, $P = 7.8 \times 10^{-67}$; Extended Data Fig. 4b). Therefore, these results collectively support the *FAM193B*-CCG repeat as an AD risk factor in the European population.

Although the *FAM193B*-CCG repeat is associated with AD in both the Chinese and European populations, its effect size is smaller in the

European population (Fig. 4f,g). To understand the genetic basis of this ethnic difference, we performed local ancestry analysis with the inclusion of additional samples of mixed European ancestry from the UK Biobank (Extended Data Fig. 5a). Interestingly, the *FAM193B*-CCG repeat positively interacted with East Asian local ancestry within its 50–150-kilobase upstream region in modulating AD risk ($\beta = 0.170$, Bonferroni-adjusted $P < 0.05$; Extended Data Fig. 5b). This interaction likely accounts for the stronger AD risk effect observed in the Chinese population. Thus, these findings establish the *FAM193B*-CCG repeat as an AD risk variant in both the Chinese and European populations, with greater disease penetrance in the Chinese population.

The *FAM193B*-CCG repeat mediates the pleiotropic association between *FAM193B* expression and AD

Given that the AD risk *FAM193B*-CCG repeat resides in the *FAM193B* promoter (Fig. 4a), it may confer disease risk by modulating *FAM193B* expression. We found that *FAM193B* was expressed across all human tissues (transcripts per million (TPM) > 10, Genotype-Tissue Expression (GTEx) dataset; Supplementary Table 17)³⁰. Moreover, single-cell transcriptomic data in human blood and brain tissues illustrate that *FAM193B* exhibited similar expression levels across cell types within each tissue and across multiple brain regions (Supplementary Figs. 4 and 5), suggesting low cell type specificity of *FAM193B* expression. To investigate the genetic regulation of *FAM193B*, we performed expression quantitative trait loci (eQTL) analysis of *FAM193B* using the blood bulk RNA sequencing (RNA-seq) data of the Hong Kong Chinese cohort³¹. Accordingly, the *FAM193B*-CCG repeat was one of the top eQTLs associated with increasing *FAM193B* expression ($\beta = 0.026$, $P = 9.1 \times 10^{-29}$, additive model; Fig. 5a and Extended Data Fig. 6). Interestingly, blood expression of *FAM193B* was higher in patients with AD than NCs ($P = 2.0 \times 10^{-15}$; Fig. 5b), suggesting that *FAM193B* expression is associated with AD.

Next, we investigated the regulatory role of the *FAM193B*-CCG repeat on *FAM193B* gene expression in the human brain. The *FAM193B*-CCG repeat was associated with increased *FAM193B* expression in the brain as reported by an eQTL study ($P = 1.3 \times 10^{-13}$; ROSMAP dataset; Extended Data Fig. 6)³². To further investigate its regulatory role among different brain cell types, we analyzed the single-nucleus RNA-seq data of postmortem human brain tissues obtained from the UK Brain Bank Network (UKBBN)³³. *FAM193B*-CCG was consistently associated with increased *FAM193B* expression across all cell subpopulations commonly detected in the dataset (Fig. 5c, Extended Data Fig. 7 and Supplementary Tables 18 and 19). Notably, it was associated with increased *FAM193B* expression in neurons ($\beta = 0.003$, $P = 3.1 \times 10^{-16}$), astrocytes ($\beta = 0.003$, $P = 2.3 \times 10^{-9}$) and oligodendrocytes ($\beta = 0.003$; $P = 2.9 \times 10^{-25}$) at genome-wide significance and in oligodendrocyte

progenitor cells ($\beta = 0.003$, $P = 6.4 \times 10^{-6}$) at suggestive significance (Fig. 5c and Supplementary Table 18). Interestingly, *FAM193B* gene expression was higher in neurons, astrocytes and oligodendrocytes in patients with AD than in NCs (adjusted $P \leq 1.5 \times 10^{-2}$; Fig. 5d and Supplementary Table 18), suggesting that, in these cell types, *FAM193B* could be involved in AD pathogenesis.

The associations of *FAM193B* expression with AD and the AD risk *FAM193B*-CCG repeat implicate *FAM193B* expression in AD pathogenesis. To examine their causal relationship, we performed summary data-based Mendelian randomization (SMR) analysis. Both the single-variant SMR test and the heterogeneity in dependent instruments (HEIDI) test indicated that the *FAM193B*-CCG repeat is a shared causal variant of AD and *FAM193B* expression in both blood (adjusted $P_{\text{SMR}} = 5.5 \times 10^{-5}$, $P_{\text{HEIDI}} = 0.239$) and the brain (adjusted $P_{\text{SMR}} = 2.8 \times 10^{-2}$, $P_{\text{HEIDI}} = 0.259$) (Fig. 5e and Supplementary Table 20). To validate the pleiotropic association between *FAM193B* expression in different tissues and AD risk, we performed a multivariate SMR (SMR-multi) test. Only the pleiotropic association between brain *FAM193B* and AD remained significant after multiple test correction (adjusted $P_{\text{SMR-multi}} = 1.7 \times 10^{-3}$; Fig. 5e,f), whereas the test for blood *FAM193B* was nominally significant ($P_{\text{SMR-multi}} = 2.3 \times 10^{-2}$; Fig. 5e), highlighting that the regulation of brain *FAM193B* expression is more important for AD. To determine which cell type(s) account for this pleiotropic association in the brain, we conducted SMR analysis with cell-type-specific *cis*-eQTL results (Supplementary Data 1). Notably, neuronal *FAM193B* expression passed all SMR tests (adjusted $P_{\text{SMR}} = 2.6 \times 10^{-2}$, adjusted $P_{\text{SMR-multi}} = 1.0 \times 10^{-2}$, $P_{\text{HEIDI}} = 0.055$; Fig. 5e,g). No other cell types showed sufficient genome-wide significant eQTLs for the SMR test (that is, microglia and oligodendrocyte progenitor cells) or the SMR-multi test (that is, astrocytes and oligodendrocytes) (Fig. 5e and Supplementary Table 20). Therefore, these results suggest that the *FAM193B*-CCG repeat is the causal variant mediating the pleiotropic association between neuronal *FAM193B* expression and AD risk.

To investigate how the *FAM193B*-CCG repeat confers causal effect on *FAM193B* gene expression, we analyzed the epigenetic profile and transcription factor binding in this CCG repeat region. The *FAM193B*-CCG-harboring region exhibited a promoter-like signature in brain tissues as indicated by strong H3K4me3 and H3K27ac signals in the frontal cortex and high chromatin accessibility across all brain cell types (Fig. 5h). In silico analysis further identified MAZ as a candidate transcription factor that binds to the *FAM193B*-CCG repeat region (Fig. 5h). Interestingly, MAZ knockdown decreased *FAM193B* expression in K562 cells (TPM decreasing from 5.2 to 2.8; Fig. 5i), supporting the role of MAZ in regulating *FAM193B* expression. To investigate whether *FAM193B*-CCG repeat length modulates the binding of *FAM193B* promoter with MAZ, we performed an electrophoretic

Fig. 5 | The *FAM193B*-CCG repeat mediates the pleiotropic association between *FAM193B* expression and AD. **a**, Association between the *FAM193B*-CCG repeat and blood *FAM193B* expression level. $n = 1,260$. Linear regression under an additive model. **b**, Association between AD and blood *FAM193B* gene expression. $n = 481$ patients with AD and $n = 566$ NCs. Robust linear regression. In **a** and **b**, box plots show medians, interquartile ranges, 10th and 90th percentiles and all individual values. **c**, Associations between the *FAM193B*-CCG repeat and *FAM193B* expression in different brain cell types. $n = 101$. Hurdle model regression under an additive model. *** $P = 6.4 \times 10^{-6}$, ### $P < 5 \times 10^{-8}$. Data are presented as effect sizes with 95% CIs. **d**, Cell-type-specific changes of brain *FAM193B* gene expression in patients with AD compared to NCs. $n = 58$ patients with AD and $n = 44$ NCs. Hurdle model regression. *Bonferroni-adjusted $P = 0.012$, ***Bonferroni-adjusted $P < 1 \times 10^{-14}$. Data are presented as effect size with 95% CI. **e**, SMR analysis of the causal effect of the *FAM193B*-CCG repeat on *FAM193B* expression and AD. **f,g**, SMR-multi plots showing the pleiotropic association between AD and *FAM193B* expression in the brain (**f**) and neurons (**g**). $n = 551$, $n = 101$ and $n = 25,670$ for brain eQTLs, neuronal eQTLs and AD association, respectively. The top *cis*-eQTL is colored red; the colors of other dots represent

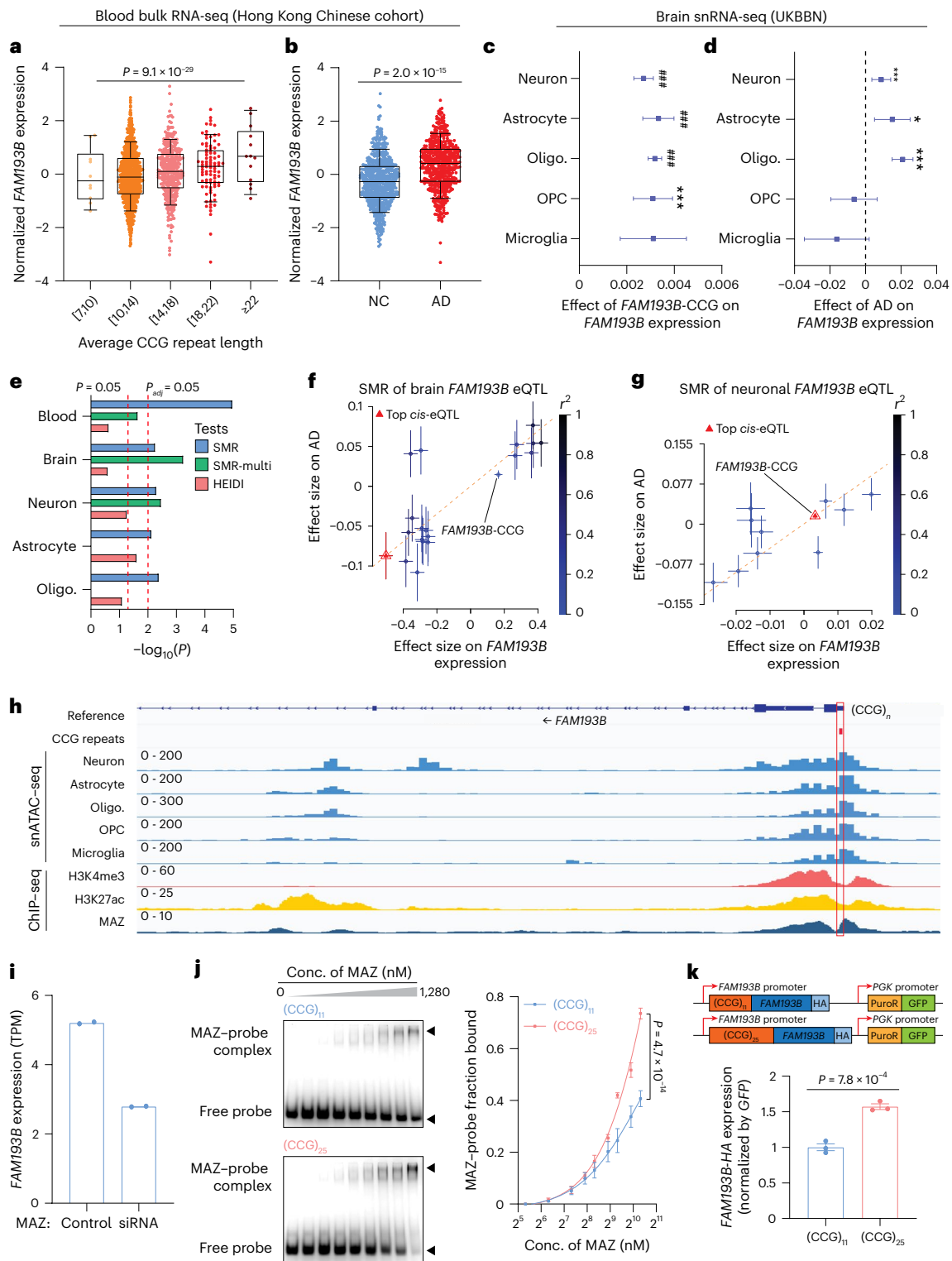
their linkage disequilibrium (as measured by r^2) with the top *cis*-eQTL. Data are presented as effect size $\pm$ s.e. **h**, Epigenetic profile and transcription factor binding in the *FAM193B* locus. The single-nucleus assay for transposase-accessible chromatin using sequencing (snATAC-seq) signals represent $-\log_{10}(P)$. The chromatin immunoprecipitation followed by sequencing (ChIP-seq) signals represent the fold change compared to the input control. **i**, Changes in *FAM193B* expression after MAZ knockdown in K562 cells. $n = 2$ biological replicates per group. Data are presented as mean $\pm$ s.e.m. **j**, Electrophoretic mobility shift assay comparing the binding between MAZ and *FAM193B*-CCG repeats with 11 and 25 CCG repeats. $n = 3$ biological replicates per group. Extra sum-of-squares *F*-test. Data are presented as mean $\pm$ s.e.m. **k**, Expression levels of *FAM193B*-HA mRNA driven by *FAM193B* promoters containing 11 or 25 CCG repeats in HEK293T cells. *FAM193B*-HA mRNA levels were normalized to *GFP*. $n = 3$ biological replicates per group. Unpaired two-tailed *t*-test. Data are presented as mean $\pm$ s.e.m. Conc., concentration; Oligo., oligodendrocyte; OPC, oligodendrocyte progenitor cell; P_{adj} , Bonferroni-corrected *P* value; siRNA, small interfering RNA; snRNA-seq, single-nucleus RNA sequencing.

mobility shift assay using DNA probes with varying CCG repeats. MAZ bound 25 CCG repeats more strongly than 11 repeats ($P = 4.7 \times 10^{-14}$; Fig. 5j), supporting a regulatory effect of the *FAM193B*-CCG repeat on *FAM193B* expression. To experimentally validate this effect, we compared *FAM193B* expression driven by *FAM193B* promoters containing different CCG repeat lengths in HEK293T cells. Promoters with 25 CCG repeats induced higher *FAM193B* gene expression than those with 11 repeats ($P = 7.8 \times 10^{-4}$; Fig. 5k and Supplementary Fig. 6). Taken together, both statistical analysis and experimental validation show that the *FAM193B*-CCG repeat is a causal variant regulating *FAM193B*

expression and mediating the pleiotropic association between *FAM193B* expression and AD.

Effects of the *FAM193B*-CCG repeat on neurodegeneration-related pathways

To understand how the *FAM193B*-CCG repeat contributes to AD risk, we examined its effects on molecular and cellular phenotypes in postmortem human brains using single-nucleus RNA-seq data. Interestingly, among all brain cell types, the *FAM193B*-CCG repeat was associated with the largest transcriptomic changes in neurons (analysis in NCs;



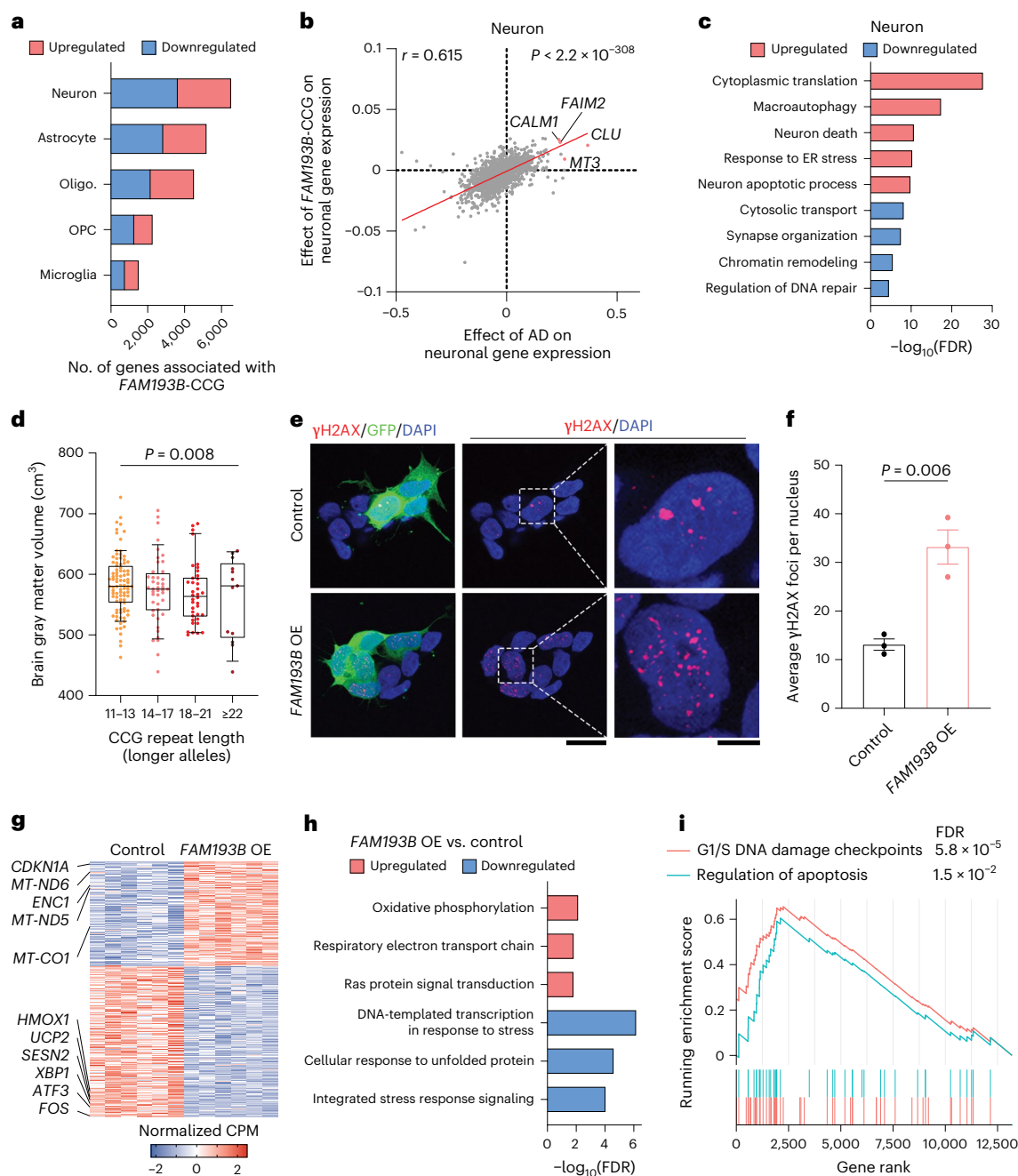


Fig. 6 | Effects of *FAM193B* on neurodegeneration-related pathways. a, Number of differentially expressed genes associated with the *FAM193B*-CCG repeat in different brain cell types. Genes with $\text{FDR} < 0.05$ were considered differentially expressed genes. $n = 44$ non-AD controls (NCs). **b**, Correlation analysis between the effects of AD and the *FAM193B*-CCG repeat on transcriptomic changes in neurons. $n = 44$ NCs for association analysis with the *FAM193B*-CCG repeat. $n = 58$ patients with AD and $n = 44$ NCs for association analysis with AD. Two-tailed correlation test. r , Pearson's correlation coefficient. **c**, Gene Ontology analysis of differentially expressed genes associated with the *FAM193B*-CCG repeat in neurons. **d**, Association between the *FAM193B*-CCG repeat and brain gray matter volume in the Hong Kong Chinese cohort. Robust linear regression under a dominant model. $n = 186$ non-AD individuals. Box plots show medians, interquartile ranges, 10th and 90th percentiles and all individual values. **e, f**, Effects of *FAM193B* overexpression (OE) on DNA damage in IMR-32 cells.

e, Representative images showing γH2AX foci in IMR-32 cells transfected with the *FAM193B* plasmid (*FAM193B* OE) and empty vector control. Cells were co-transfected with pmaxGFP to identify transfected cells. Nuclei were counterstained with DAPI. Scale bars, 20 μm (left) and 5 μm (right). **f**, Quantification of the average number of nuclear γH2AX foci in GFP⁺ cells. $n = 3$ biological replicates per group. Unpaired two-tailed t -test. Data are presented as mean $\pm$ s.e.m. **g**, Heatmap showing the transcriptomic profile of IMR-32 cells transfected with the *FAM193B* plasmid (*FAM193B* OE) and empty vector control. Only significant differentially expressed genes (that is, $\text{FDR} < 0.05$) are shown. $n = 6$ biological replicates per group. **h**, Gene Ontology analysis of differentially expressed genes in *FAM193B*-overexpressing IMR-32 cells. **i**, Gene set enrichment analysis showing the upregulation of DNA damage-related pathways in *FAM193B*-overexpressing IMR-32 cells. Genes are ranked by Wald statistics. CPM, counts per million; Oligo., oligodendrocyte; OPC, oligodendrocyte progenitor cell.

Fig. 6a). Notably, the association between neuronal gene expression and the *FAM193B*-CCG repeat was correlated with the association between neuronal gene expression and AD ($r = 0.615$, $P < 2.2 \times 10^{-308}$; Fig. 6b), indicating that the *FAM193B*-CCG repeat may affect AD-related

pathways. Gene Ontology analysis illustrated that the genes positively associated with the *FAM193B*-CCG repeat in neurons were enriched in cell death-related pathways, such as macroautophagy, neuron death and neuron apoptotic process ($\text{FDR} < 0.05$; Fig. 6c).

To investigate the effect of the *FAM193B*-CCG repeat on neurodegeneration, we examined its association with brain gray matter volumes measured by magnetic resonance imaging in the Hong Kong Chinese cohort. Accordingly, longer *FAM193B*-CCG repeats were associated with decreased gray matter volume ($\beta = -0.015$, s.e. = 0.006, $P = 8.3 \times 10^{-3}$; Fig. 6d), suggesting more severe neurodegeneration. To investigate this functional consequence, we examined the effects of *FAM193B* expression on neuronal death-related pathways using a neuroblastoma IMR-32 cell line. *FAM193B* overexpression resulted in increased γ H2AX foci formation ($P = 0.006$; Fig. 6e,f and Supplementary Fig. 7), a molecular marker of DNA damage³⁴. Regarding the molecular pathways regulated by *FAM193B*, RNA-seq analysis identified 270 differentially expressed genes (FDR < 0.05) associated with *FAM193B* overexpression in IMR-32 cells (Fig. 6g). Specifically, upregulated genes were enriched in pathways related to oxidative phosphorylation (FDR = 7.2×10^{-3}) and Ras protein signal transduction (FDR = 1.5×10^{-2}); downregulated genes were enriched in stress response pathways (FDR < 6.0×10^{-5}) (Fig. 6h). These findings suggest that *FAM193B* overexpression results in stress response dysregulation and affects mitochondria-mediated energy production and reactive oxygen species production pathways in neuronal cells. Moreover, examination of global transcriptomic changes by gene set enrichment analysis revealed significant upregulation of genes involved in G1/S DNA damage checkpoints (FDR = 5.8×10^{-5}) and regulation of apoptosis (FDR = 1.5×10^{-2}) in *FAM193B*-overexpressing cells (Fig. 6i). Thus, these findings suggest the role of *FAM193B* in modulating DNA damage and promoting neuronal death.

Discussion

This study highlights the contribution of common SVs to AD and identifies the *FAM193B*-CCG repeat polymorphism as an AD risk genetic factor. By leveraging LRS and SRS data, we identified and genotyped high-confidence SVs in a Chinese AD cohort. As a proof of concept, PRS analysis with 10-fold cross-validation shows that the polygenic effects of common SVs account for additional AD risk beyond that explained by common SNPs. Among these, the *FAM193B*-CCG repeat emerged as a common AD risk variant mediating the pleiotropic association between brain *FAM193B* expression and AD. Analysis of human data and cellular assays consistently showed that elevated *FAM193B* expression results in increased DNA damage and activation of cell death-related pathways in neurons. Therefore, these results establish the *FAM193B*-CCG repeat as a causal variant contributing to AD risk and neurodegeneration, underscoring the importance of common SVs in genetic studies of AD.

Despite the identification of numerous AD risk SNPs³, the estimated SNP heritability of AD is notably lower than that observed in twin studies². In addition, SVs are increasingly recognized as key genetic components that potentially explain this missing heritability of AD¹⁰. Our PRS analysis revealed that common SVs contribute to additional disease risk beyond that explained by common SNPs. Nevertheless, larger discovery cohorts and independent replication analysis are required to precisely estimate SNP-based or SV-based heritability for AD. Interestingly, we found that approximately 18% of common SVs were not tagged by SNPs, particularly those located within tandem repeats (Supplementary Fig. 8). These untagged SVs may represent a substantial portion of the genetic risk of AD poorly estimated by common SNPs. Therefore, including SVs in PRS models could provide a more comprehensive estimate of global genetic risk, thereby improving risk prediction for human diseases. Furthermore, common SVs can exert larger effect sizes on human phenotypes than common SNPs¹⁴. For instance, 5.9% of Chinese individuals harbored at least 21 *FAM193B*-CCG repeats, which strongly increases AD risk (OR = 2.89, 95% CI: 1.98–4.22; Supplementary Table 15). Hence, the strong AD risk effects of common SVs present opportunities for developing genetic biomarkers of AD. Therefore, including SV analysis in AD genetic studies would extend understanding of the genetic architecture of AD and facilitate lifetime disease risk prediction in clinical settings.

Tandem repeat expansions are known causal factors of specific neurological disorders, such as Huntington's disease³⁵. Although a recent study shows that patients with AD have a higher polygenic burden of rare tandem repeat expansions than NCs¹³, we identified a common tandem repeat polymorphism in the *FAM193B* promoter that is associated with increased AD risk, suggesting that common tandem repeat polymorphisms also contribute to AD risk. Tandem repeats in regulatory regions can regulate gene expression by modulating transcription factor binding to regulatory elements³⁶. Our in vitro assays show that the *FAM193B*-CCG repeat increases the transcription factor binding affinity of the *FAM193B* promoter and upregulates *FAM193B* expression, suggesting that the common tandem repeats within regulatory regions are important genetic factors for modulating gene expression and AD risk. Further genome-wide analysis of common tandem repeats could elucidate new genetic mechanisms underlying AD pathogenesis.

FAM193B is a protein-coding gene broadly expressed across multiple tissues and cell types (Supplementary Figs. 4 and 5 and Supplementary Table 17). Although its biological function remains unclear, recent LRS studies link rare *FAM193B*-CCG repeat expansions to rare diseases^{37,38}, suggesting that *FAM193B* is involved in important biological processes. Here, we identified *FAM193B* as an AD risk locus, with the common *FAM193B*-CCG repeat polymorphism exerting a causal effect. Interestingly, a recent AD GWAS in the European population reports rs335480 as a genome-wide significant AD risk SNP in the *FAM193B* locus²⁹. Our linkage disequilibrium analysis, conditional analysis and SMR analysis collectively suggest that the *FAM193B*-CCG repeat is the causal variant tagged by rs335480 in the European population (Fig. 5e–g, Extended Data Fig. 4 and Supplementary Table 14). Although the *FAM193B*-CCG repeat is associated with increased *FAM193B* expression across multiple tissues and cell types, our findings highlight that neuronal *FAM193B* plays an important role in AD pathogenesis (Figs. 5e,g and 6a). This is supported by cellular experiments showing that increased *FAM193B* expression in cultured neuronal cells promotes DNA damage and activates cell death-related pathways (Fig. 6e–i). Interestingly, *FAM193B* protein directly interacts with GSK3 α and GSK3 β ³⁹, the glycogen synthase kinases whose dysregulation results in increased β -amyloid production and tau hyperphosphorylation⁴⁰. This provides hints about the implications of *FAM193B* in known AD pathogenic pathways. Therefore, functional studies are required to investigate the molecular mechanisms underlying the roles of *FAM193B* in neurodegeneration and AD pathogenesis.

In summary, by investigating common SVs, we identify *FAM193B* as an AD susceptibility locus, revealing a common CCG repeat polymorphism as the underlying causal variant. Including SVs in large-scale GWASs will help resolve missing heritability, advance mechanistic insights and provide new candidate therapeutic targets.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41593-026-02481-z>.

References

- 2023 Alzheimer's disease facts and figures. *Alzheimers Dement.* **19**, 1598–1695 (2023).
- Andrews, S. J. et al. The complex genetic architecture of Alzheimer's disease: novel insights and future directions. *EBioMedicine* **90**, 104511 (2023).
- Bellenguez, C. et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* **54**, 412–436 <https://doi.org/10.1038/s41588-022-01024-z> (2022).

4. Farrer, L. A. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. *JAMA* **278**, 1349–1356 (1997).
5. Leonenko, G. et al. Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nat. Commun.* **12**, 4506 (2021).
6. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
7. Pang, A. W. et al. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* **11**, R52 (2010).
8. Ming, C. et al. Whole genome sequencing-based copy number variations reveal novel pathways and targets in Alzheimer's disease. *Alzheimers Dement.* **18**, 1846–1867 (2022).
9. Hooli, B. V. et al. Rare autosomal copy number variations in early-onset familial Alzheimer's disease. *Mol. Psychiatry* **19**, 676–681 (2014).
10. Wang, H., Wang, L. S., Schellenberg, G. & Lee, W. P. The role of structural variations in Alzheimer's disease and other neurodegenerative diseases. *Front. Aging Neurosci.* **14**, 1073905 (2023).
11. Rovelet-Lecrux, A. et al. *APP* locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat. Genet.* **38**, 24–26 (2005).
12. Wang, H. et al. Structural variation detection and association analysis of whole-genome-sequence data from 16,543 Alzheimer's disease sequencing project subjects. *Alzheimers Dement* **21**, e70277 (2025).
13. Guo, M. H., Lee, W.-P., Vardarajan, B., Schellenberg, G. D. & Phillips-Cremins, J. E. Polygenic burden of short tandem repeat expansions promotes risk for Alzheimer's disease. *Nat. Commun.* **16**, 1126 (2025).
14. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
15. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
16. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
17. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).
18. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
19. Chen, S. et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
20. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
21. Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
22. Gabel, H. W. et al. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature* **522**, 89–93 (2015).
23. Kleinert, P. & Kircher, M. A framework to score the effects of structural variants in health and disease. *Genome Res.* **32**, 766–777 (2022).
24. Willems, T. et al. Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* **14**, 590–592 (2017).
25. Mousavi, N., Shleizer-Burko, S., Yanicky, R. & Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).
26. Dolzhenko, E. et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).
27. Zhou, X. et al. Identification of genetic risk factors in the Chinese population implicates a role of immune system in Alzheimer's disease pathogenesis. *Proc. Natl Acad. Sci. USA* **115**, 1697–1706 (2018).
28. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
29. Lambert, J. C. et al. Consensus meta-analysis of genome-wide association studies for Alzheimer's disease and related dementias. *Nat. Genet.* **58**, 1214–1225 (2026).
30. Aguet, F. et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
31. Zhong, H. et al. Using blood transcriptome analysis for Alzheimer's disease diagnosis and patient stratification. *Alzheimers Dement.* **20**, 2469–2484 <https://doi.org/10.1002/ALZ.13691> (2024).
32. Cui, Y. et al. Multi-omic quantitative trait loci link tandem repeat size variation to gene regulation in human brain. *Nat. Genet.* **57**, 369–378 (2025).
33. Lau, S.-F., Cao, H., Fu, A. K. Y. & Ip, N. Y. Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **117**, 25800–25809 (2020).
34. Mah, L. J., El-Osta, A. & Karagiannis, T. C. γH2AX: a sensitive molecular marker of DNA damage and repair. *Leukemia* **24**, 679–686 (2010).
35. Penney, J. B., Vonsattel, J. P., MacDonald, M. E., Gusella, J. F. & Myers, R. H. CAG repeat number governs the development rate of pathology in Huntington's disease. *Ann. Neurol.* **41**, 689–692 (1997).
36. Horton, C. A. et al. Short tandem repeats bind transcription factors to tune eukaryotic gene expression. *Science* **381**, eadd1250 (2023).
37. Jensen, T. D. et al. Integration of transcriptomics and long-read genomics prioritizes structural variants in rare disease. *Genome Res.* **35**, 914 (2025).
38. Danzi, M. C. et al. Detailed tandem repeat allele profiling in 1,027 long-read genomes reveals genome-wide patterns of pathogenicity. Preprint at *bioRxiv* <https://doi.org/10.1101/2025.01.06.631535> (2025).
39. Cormier, K. W., Larsen, B., Gingras, A. C. & Woodgett, J. R. Interactomes of glycogen synthase kinase-3 isoforms. *J. Proteome Res.* **22**, 977–989 (2023).
40. Hooper, C., Killick, R. & Lovestone, S. The GSK3 hypothesis of Alzheimer's disease. *J. Neurochem.* **104**, 1433–1439 (2008).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

¹Division of Life Science, State Key Laboratory of Nervous System Disorders, Daniel and Mayce Yu Molecular Neuroscience Center, The Hong Kong University of Science and Technology, Hong Kong Special Administrative Region, China. ²InnoHK Hong Kong Center for Neurodegenerative Diseases, Hong Kong Special Administrative Region, China. ³Guangdong Provincial Key Laboratory of Brain Science, Disease and Drug Development, Shenzhen–Hong Kong Institute of Brain Science, HKUST Shenzhen Research Institute, Shenzhen, China. ⁴SIAT–HKUST Joint Laboratory for Brain Science, Shenzhen, China. ⁵The Brain Cognition and Brain Disease Institute, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences; Shenzhen–Hong Kong Institute of Brain Science—Shenzhen Fundamental Research Institutions, Shenzhen, China. ⁶Center for Stem Cell Research, HKUST–Nan Fung Life Sciences Joint Laboratory, The Hong Kong University of Science and Technology, Hong Kong Special Administrative Region, China. ⁷Therese Pei Fong Chow Research Centre for Prevention of Dementia, Division of Geriatrics, Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China. ⁸Department of Neurodegenerative Disease, Queen Square Institute of Neurology, University College London, London, UK. ⁹UK Dementia Research Institute, University College London, London, UK. ✉e-mail: boamy@ust.hk; boip@ust.hk

Methods

Sample recruitment and data collection

For the Hong Kong Chinese cohort, we recruited 2,333 participants from the Specialist Outpatient Clinics of the Department of Medicine and Therapeutics at the Prince of Wales Hospital, Haven of Hope Christian Service and Community CareAge Foundation. All participants underwent the Montreal Cognitive Assessment (MoCA)⁴¹, and patients with AD were diagnosed according to the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition: DSM-5* (ref. 42). This study was approved by the Clinical Research & Ethics Committees of the Joint Chinese University of Hong Kong–New Territories East Cluster for the Prince of Wales Hospital as well as the Institutional Review Board, Hospital Authority (CREC ref. nos. 2015.461 and 2021.651) and the Human Participants Research Panel of the Hong Kong University of Science and Technology (CRP#180, HREP-2023-0179). All participants provided written informed consent for both study enrollment and sample collection.

We also retrieved additional human whole-genome sequencing, bulk RNA-seq, single-nucleus RNA-seq and brain magnetic resonance imaging data from previous studies and publicly available databases. The description of those data can be found in the Supplementary Methods.

Generation of long-read whole-genome sequencing data

We collected the PacBio continuous long-read data of 11 Chinese individuals, including nine publicly available samples (Supplementary Table 1) and two in-house participants recruited from the Specialist Outpatient Department at the Prince of Wales Hospital. Among these 11 individuals, six are parent–child trios (that is, HG00512, HG00513 and HG00514 as well as HG005, HG006 and HG007), and the rest are unrelated.

For the in-house LRS data, we performed library preparation according to 'Procedure & Checklist > 20kb Template Preparation Using BluePippin Size-Selection System (15–20 kb Cutoff) for Sequel Systems' from PacBio using a PacBio Template Prep Kit 1.0 (PacBio, 100-259-100). In brief, we sheared 10 µg of DNA to the desired size using a g-TUBE (Covaris, 520079). After purification with AMPure PB beads (PacBio, 100-265-900), we removed single-stranded ends by ExoVII treatment and repaired DNA using damage repair mix and end repair mix. After blunt-end ligation, we size-selected the DNA using a BluePippin size-selection system, annealed it with primers and bound it with polymerase (PacBio, 100-619-300). We conducted all purification steps according to the manufacturer's protocol. We then loaded the library using Magbeads (PacBio, 100-125-900) and sequenced it on a Sequel 1.0 system using a Sequel Sequencing Kit 1.2 (PacBio, 100-902-200).

We converted the raw LRS data generated by the PacBio RS II platform into subreads using PacBio's bax2bam (version 0.0.8) with the default settings⁴³. After combining individuals' subreads, we selected the representative sequences of subreads using PacBio's ccs (version 3.4 for PacBio RS II, version 4 for PacBio Sequel)⁴³ with the following parameters: --maxLength 10Mbp, --minPasses 0 and --noPolish. We evaluated the quality of the representative long-read sequences by using LongQC's sampleqc function⁴⁴.

Generation of short-read whole-genome sequencing data

We performed DNA extraction from 400 µl of blood cell pellets using a QIAamp DNA Blood Mini Kit (Qiagen) on a QIAcube Connect machine (Qiagen). We eluted the extracted DNA twice using 50 µl and 60 µl of AE buffer, respectively, which yielded approximately 100 µl of DNA samples. We evaluated the concentration, A260/280 and A260/230 of the DNA samples using a BioDrop Duo+ spectrophotometer (Biochrom). Novogene performed library preparation and high-coverage (average: 35× after deduplication), 150-base-pair, paired-end, whole-genome sequencing using an Illumina NovaSeq 6000 System.

SV calling using LRS data

We developed a workflow (available at <https://github.com/Han-Cao/HarmoniSV>) to perform joint SV calling across samples and methods, which includes SV discovery, SV merging, SV re-genotyping and SV filtering.

SV discovery. We aligned clean LRS data to the GRCh38 human reference genome using minimap2 and NGMLR^{16,20}. We then subjected the alignments from each aligner to three different SV callers—Sniffles2 (ref. 16), cuteSV¹⁸ and SVIM¹⁷—using the default parameters optimized for PacBio continuous long-read data. We detected duplicate SV calls using Jasmine (version 1.1.5) with the following parameters: --allow_intrasample --nonlinear_dist max_dist = 200 (ref. 45). For duplicate SVs, we kept only the one with most supporting reads.

SV merging. To identify non-redundant SV calls from different individuals and methods, we merged the identified SVs using Jasmine with the --ignore_strand parameter⁴⁵. Among the merged SVs, we determined the representative SV by selecting the SV with the highest frequency; alternatively, if multiple SVs shared the same maximum frequency, we selected the one with the shortest Euclidean distance to the average start position and length as follows: $\arg \min \{ \sqrt{(s - s_{\text{mean}})^2 + (l - l_{\text{mean}})^2} \}$, where s denotes the start position and l denotes SV length.

SV re-genotyping. We performed re-genotyping analysis of merged SVs using Sniffles2 and cuteSV with the default parameters^{16,18}. We determined the genotypes of the SVs based on the average alternative allele fraction of the re-genotyping results. To achieve sensitive SV discovery, we included only non-reference calls (that is, allele fraction > 0.2) to calculate the average alternative allele fraction.

SV filtering. To obtain high-confidence SVs, we trained random forest models separately for insertions and deletions using the scikit-learn Python package⁴⁶. We overlapped the identified SVs with published SV datasets (Supplementary Table 5) to obtain training sites. Specifically, we defined positive SVs as those existing in at least four published SV datasets and negative SVs as novel SVs that were not detected by all methods or any SV with fewer than four supporting reads. We subjected multiple features of SV calls to the random forest model, including SV length, alternative allele fraction, the standard deviation of allele fractions and the depths of individual SV calling methods. We performed model selection by 10-fold cross-validation. To evaluate model performance, we performed the aforementioned SV calling methods on HG002 and benchmarked the identified SVs with HG002 ground truth SVs in the tier 1 regions using the following commands: `truvari --pctsim 0 -r 1000` (refs. 47,48). We calculated AUROCs using the pROC R package⁴⁹. To retain an FDR ≤ 5% of the SV calling results, we filtered insertions and deletions according to random forest score cutoffs of >0.779 and >0.135, respectively. We measured the Mendelian error rate in the two family trios by using the BCFtools mendelian plugin⁵⁰.

Variant calling and genotyping using SRS data

We subjected the raw sequencing data to fastp for quality control, read trimming and filtration⁵¹. We used BWA-mem to align clean reads to the GRCh38 human reference genome and then subjected the reads to Genome Analysis Toolkit (GATK) for duplicate removal and base quality score recalibration^{52,53}. We performed variant calling by using GATK's germline joint calling workflow and further filtered identified SNPs and small insertions and deletions by variant quality score recalibration with sensitivity cutoffs of 95% and 90% for SNPs and for insertions and deletions, respectively⁵³. We used Paragraph (version 2.4a) to perform SV genotyping of SVs identified from the LRS data¹⁹.

After variant calling and genotyping, we performed conventional quality control analysis. All participants showed consistent sex

according to their genotypes and phenotypes. We calculated relatedness by performing KING analysis using PLINK2 (ref. 54) and applied a kinship coefficient cutoff of 0.08 to remove second-degree relatives ($n = 194$). We calculated principal components and removed genetic outliers ($n = 16$) according to the instruction manual of the bigsnpr R package⁵⁵. After sample-level quality control, we kept clean SNPs and SVs that were autosomal, biallelic and non-singleton and with Hardy–Weinberg equilibrium P values greater than 1×10^{-6} .

To further remove SVs with low genotyping accuracy, we conducted the same genotyping analysis using the 35× SRS data of five LRS samples. We then compared the concordance between LRS and SRS data and kept only the SVs that had 100% concordance with LRS genotypes. Finally, we genotyped 17,294 high-confidence SVs in 2,123 unrelated individuals, including 6,792 common SVs with MAF > 0.05 for genome-wide association analysis and PRS analysis.

Genotyping of the *FAM193B*-CCG repeat

We used HipSTR (version 0.6.2), GangSTR (version 2.5.0) and ExpansionHunter (version 5.0.0) to genotype the *FAM193B*-CCG repeat^{24–26}. For HipSTR, we used the default parameters to jointly genotype all samples. We ran GangSTR (with the parameters --grid-threshold 250 --readlength 150) and ExpansionHunter (with the default parameters) on individual samples and used MergeSTR (version 6.1.0) to merge all samples of the same method into a population call set⁵⁶. We then used EnsembleTR (version 1.0.0) with the default parameters to combine the genotyping results from different methods. We retained high-confidence *FAM193B*-CCG repeat genotypes by keeping genotypes with an EnsembleTR score ≥ 0.4 . The difference in analytical sample size between the *FAM193B*-CCG repeat and other genetic variants is equal to the number of samples with EnsembleTR score < 0.4. Although the GRCh38 reference genome has 13 CCG repeats in the genomic region chr5:177,554,492–177,554,530, we found that most samples had a C base deletion at chr5:177,554,495. Therefore, we only quantified CCG repeats after chr5:177,554,497 for all analyses, which corresponds to a reference length of 11 CCG repeats.

Functional annotation and enrichment of SVs

We used Ensembl Variant Effect Predictor (version 105) to annotate SVs according to their overlap with genomic features⁵⁷. We used CADD-SV scores to analyze the deleteriousness of SVs according to their sequence and flanking genomic features²³. To annotate SVs with known disease risk loci, we obtained the GWAS results from the NHGRI-EBI GWAS Catalog (accessed on 3 November 2022)²¹. We defined the disease risk loci as the 10-kilobase flanking region of reported genes mapped by significant disease risk variants ($P < 5 \times 10^{-8}$) and annotated their overlap with SVs using BEDTools⁵⁸.

To explore the overrepresentation or underrepresentation of SVs in genomic features, we conducted enrichment analysis by calculating the following statistic:

$$OR = \frac{\text{overlap bp with feature/feature size}}{\text{overlap bp with reference/reference size}}$$

where the reference sizes used to analyze functional genomic elements and gene biotypes are the lengths of the human genome and all human genes, respectively. We calculated effect sizes and statistical significance by the bootstrap method with 1,000 replicates using the 'boot' and 'boot.pval' packages in R^{59,60}.

Association analysis between genetic variants and AD

The statistical models and adjusted covariates of all regression analyses in this study are provided in Supplementary Table 21. Specifically, in the Chinese cohorts, we used PLINK2 to perform genome-wide logistic regression analysis of AD⁵⁴, adjusting for age, sex and the first three principal components. To further estimate the effect sizes of

FAM193B-CCG repeats of different lengths, we performed logistic regression with the dominant model to compare the AD risk of individuals harboring at least n CCG repeats ($n > 11$) against those with ≤ 11 CCG repeats. We used METASOFT to meta-analyze the association between AD and the *FAM193B*-CCG repeat in the Chinese population using a fixed-effect meta-analysis model⁶¹. Given that the *FAM193B*-CCG repeat exerts a more significant AD risk effect in the dominant model than in the additive model (Supplementary Table 13), we used the dominant model to examine the effect of the *FAM193B*-CCG repeat on AD-related endophenotypes in downstream analysis (for example, brain transcriptome and brain gray matter volume analysis).

To replicate the association between AD and the *FAM193B*-CCG repeat in the European population, we performed whole-genome regression analysis using REGENIE⁶². Specifically, in REGENIE step 1, we calculated genetic predictors using 488,696 high-quality, independent array SNPs (selected by PLINK2 --maf 0.01 --mac 200 --geno 0.01 --hwe 1e-15 --indep-pairwise 1000 100 0.9) with adjustment for age, sex, whole-genome sequencing center and the first 20 principal components. We then performed REGENIE step 2 association analysis with adjustment for the same covariates and the genetic predictors obtained from step 1.

Conditional analysis

To examine whether the *FAM193B*-CCG repeat confers AD risk effect independent of nearby SNPs, we used PLINK2 and REGENIE to perform reciprocal conditional analysis between the *FAM193B*-CCG repeat and lead AD-associated SNPs in the Hong Kong Chinese cohort and UK Biobank, respectively. Specifically, for the Hong Kong Chinese cohort, we conditioned on the lead SNP rs182133156 ($P = 5.6 \times 10^{-4}$); for the UK Biobank, we conditioned on rs335480, which is the lead SNP reported in a large-scale AD GWAS in the European population²⁹. As no SNP in the *FAM193B* locus (within 200 kilobases) passed Bonferroni correction after conditioning on the *FAM193B*-CCG repeat, we did not include additional SNPs for step-wise conditional analysis. The adjusted covariates are the same as those in the association analysis described above.

PRS analysis

To examine the contributions of common SVs to AD, we conducted clumping plus thresholding PRS analysis. We clumped all genetic variants using PLINK2 --clump-r2 0.1 --clump-kb 500 with different P value cutoffs (Supplementary Table 10)⁵⁴. We repeated 10-fold cross-validation 10 times to calculate the correlations between PRSs and AD. Specifically, we used nine folds of samples to estimate the effect size and used the remaining one fold of samples for validation. To evaluate the models, we calculated the correlations between PRSs and AD according to the increase in Nagelkerke's pseudo r^2 with the addition of PRSs to the model that only included covariates (that is, age, sex and the first three principal components). We calculated the AUROCs of the PRS models using the pROC R package⁴⁹. To construct the SNP + SV models, we further clumped selected SNPs and SVs to filter SNPs or SVs in linkage disequilibrium with each other. We fitted the SNP + SV models using the following logistic regression model:

$$AD = PRS_{SNP} + PRS_{SV} + \text{age} + \text{sex} + PC1 + PC2 + PC3$$

We then calculated PRSs of the SNP + SV models as a weighted score based on the PRSs of the SNP and SV models as follows:

$$PRS_{SNP+SV} = \beta_{SNP} \times PRS_{SNP} + \beta_{SV} \times PRS_{SV}$$

where β is the beta coefficient obtained from the above regression.

We further conducted conditional PRS analysis to control for the effect of linkage disequilibrium between SVs and SNPs for the SNP + SV models. Specifically, we performed association analysis for SNPs with

adjustment for each nearby lead SV (that is, within 1 megabase) that was included in the SNP + SV models. If an SNP had more than one nearby lead SV, we used the statistics of the conditional analysis with the largest P value to fully adjust for the effects of SVs. We then repeated the above clumping analysis on these adjusted statistics to include SNPs for the SNP_{adj} + SV models. This approach ensured that the effects of SNPs and SVs included in the SNP_{adj} + SV models are fully independent. We compared the performance among PRS models within each fold by paired two-tailed Wilcoxon signed-rank test.

To validate the PRS models, we applied the PRS models of a representative fold (that is, one with a PRS r^2 value close to the mean of the cross-validation results) to its validation set as well as additional MCI samples. We examined the correlation of PRSs derived from different models using the `cor.test` function in R. To compare the performance of different PRS models for distinguishing MCI from NC, we performed the DeLong test using `roc.test` in the `pROC` R package⁴⁹ and estimated the net reclassification improvement using the following function from the `nrncens` R package: `nrbin(cut = 0.05, updown = 'diff', niter = 0)`⁶³.

eQTL mapping of *FAM193B*

Descriptions of the transcriptomic dataset and related preprocessing analysis are provided in the Supplementary Methods. We performed *cis*-eQTL analysis of blood *FAM193B* expression using the samples with both whole-genome sequencing and bulk RNA-seq data in the Hong Kong Chinese cohort ($n = 1,315$). We conducted *cis*-eQTL analysis using the linear regression model in PLINK2 with adjustment for age, sex, the first three principal components and 60 probabilistic estimation of expression residuals (PEER) factors⁶⁴. We also obtained the brain *FAM193B* eQTL summary statistics of the ROSMAP dataset from Cui et al.³² and de Klein et al.⁶⁵ for the *FAM193B*-CCG repeat and SNPs, respectively. To further investigate the genetic regulation of *FAM193B* in different brain cell types, we performed cell-type-specific *cis*-eQTL analysis of brain *FAM193B* expression using the single-nucleus RNA-seq data of samples obtained from the UKBBN ($n = 102$). We used the hurdle model from the `MASTR` package⁶⁶ to perform regression analysis with adjustment for age, sex, postmortem duration, gene detection rate and the first three principal components. To validate candidate *cis*-eQTLs, we obtained the brain single-cell *cis*-eQTL summary statistics from the SingleBrain dataset ($n = 983$)⁶⁷. Among *cis*-eQTLs identified in the UKBBN with suggestive significance (that is, $P < 1 \times 10^{-5}$), we kept those that reached genome-wide significance (that is, $P < 5 \times 10^{-8}$) in the SingleBrain dataset and exhibited a concordant effect direction. We applied Bonferroni adjustment to correct for multiple testing on genetic variants or brain cell types.

SMR analysis

We performed SMR analysis to infer the causal effects of the *FAM193B*-CCG repeat on *FAM193B* expression and AD as well as their pleiotropic associations. For blood SMR analysis, we used the AD association summary statistics and *FAM193B* *cis*-eQTL results from the Hong Kong Chinese cohort. To avoid the SMR analysis being confounded by overlapping samples used in both the AD GWAS and *FAM193B* eQTL analysis, we excluded participants with AD from eQTL analysis to prepare summary statistics ($n = 836$) specifically for SMR analysis. For brain SMR analysis, we used the bulk *FAM193B* eQTL summary statistics of the ROSMAP dataset from Cui et al.³² and de Klein et al.⁶⁵ for the *FAM193B*-CCG repeat and nearby SNPs, respectively. We also used the single-cell *FAM193B* *cis*-eQTL results from the UKBBN samples for brain cell-type-specific SMR analysis. The AD risk effects of *FAM193B* *cis*-eQTLs in the European population were estimated in the UK Biobank dataset as described above.

We conducted a single-variant SMR test and a HEIDI test on the *FAM193B*-CCG repeat along with an SMR-multi test for all genome-wide significant *cis*-eQTLs using SMR with the following parameters:

--ld-lower-limit 0.01 --target-snp FAM193B-CCG --smr-multi⁶⁸. Given the weak linkage disequilibrium between the *FAM193B*-CCG repeat and nearby SNPs, we used a smaller --ld-lower-limit of 0.01 (compared to the default value of 0.05) to include more SNPs for the HEIDI test. We applied Bonferroni adjustment to correct for multiple SMR tests in different tissues and cell types.

Electrophoretic mobility shift assay

DNA oligonucleotides with and without Cy5 labeling (Supplementary Table 22) were synthesized by GenScript and dissolved in water. We mixed oligonucleotides in annealing buffer (10 mM Tris-HCl, 50 mM NaCl and 1 mM EDTA (pH 8.0)) to obtain a 10 μM solution of double-stranded DNA. We incubated the solution at 95 °C for 5 minutes and then cooled it to room temperature. We then dissolved recombinant MAZ protein (MyBioSource, MBS952961) in water. We mixed 0.2 pmol of annealed double-stranded DNA with different concentrations of recombinant MAZ protein and 1 ng μl^{-1} Calf Thymus DNA Solution (Thermo Fisher Scientific, 15633-019) in binding buffer (250 mM HEPES, 500 mM KCl, 20 mM MgSO_4 and 10 mM dithiothreitol (DTT) (pH 8.0)). After incubation in the dark for 2 hours, we added glycerol to the mixture to a final concentration of 5%. We then subjected the mixture to electrophoresis on 6% non-denaturing polyacrylamide gels on ice in 0.5× TBE buffer. We visualized gels using a ChemiDoc MP System (Bio-Rad). We used ImageJ (version 1.53q) to quantify the band intensities of free DNA probes and MAZ-probe complexes. We used Graph-Pad Prism (version 8.0.2) to fit binding curves with a four-parameter logistic regression model and compared the fitted models by extra sum-of-squares F -test.

Validation of the effect of the *FAM193B*-CCG repeat on *FAM193B* gene expression

We used a piggyBac-based dual-promoter expression system to evaluate the regulatory effect of the *FAM193B* CCG repeat, which was modified from PB-CMV-MCS-Puro-EGFP (Addgene, 219795; a gift from K. Nakayama). Specifically, we engineered the two constructs to contain the *FAM193B* promoter harboring either 11 or 25 CCG repeats, driving the expression of *FAM193B* cDNA (RefSeq accession number: [NM_001190946](#)) fused with an HA tag at the C terminus. A puromycin resistance gene and *GFP* were expressed from an independent *PGK* promoter to enable antibiotic selection and control of integration efficiency (see Supplementary Methods for construct sequences).

We co-transfected HEK293T cells (American Type Culture Collection (ATCC), CRL-3216) with either the 11 or the 25 CCG repeats construct together with pCMV-hyPBase⁶⁹ using EZ Trans Cell Transfection Reagent (Life-iLab, ACO4L092). Two days after transfection, we selected cells with 2 $\mu\text{g ml}^{-1}$ puromycin (Thermo Fisher Scientific, A1113803) for 3 days and maintained them in culture with 1 $\mu\text{g ml}^{-1}$ puromycin for 11 days. After selection, we sorted the top 20% of GFP-expressing cells based on GFP fluorescence intensity using a BD Influx cell sorter (BD Biosciences) and extracted total RNA using an RNeasy Plus Mini Kit (Qiagen, 74134) on a QIAcube machine. We quantified the extracted RNA samples with a BioDrop spectrophotometer. We performed cDNA synthesis with a PrimeScript RT Reagent Kit (Takara, RR037A) followed by quantitative polymerase chain reaction (qPCR) using Taq Pro Universal SYBR Master Mix (Vazyme, Q712). We detected exogenous *FAM193B*-HA transcripts using primers spanning the *FAM193B* 3' end and HA tag (forward: 5'-GTCAGAAAGTTGCTGTGAAC-3'; reverse: 5'-AGCGTAATCTGGAACATCGTAT-3'). We quantified *GFP* expression (forward: 5'-GAACCGCATCGAGCTGAA-3'; reverse: 5'-TGCTTGTGGCCATGATATAG-3') to control for integration efficiency and used *GAPDH* (forward: 5'-GCACCGTCAAGGCTGAGAAC-3'; reverse: 5'-AGGGATCTCGCTCCTGGAA-3') as the input RNA control. We calculated the relative expression of *FAM193B*-HA using the $-\Delta\Delta\text{Ct}$ method normalized to either *GAPDH* or *GFP*. We evaluated fold change differences by unpaired two-tailed t -test.

DNA damage analysis and RNA-seq in IMR-32 cells

The *FAM193B* overexpression plasmid was constructed by GenScript by synthesizing the *FAM193B* cDNA fragment (RefSeq accession number: [NM_001190946](#)) and subcloning it into the pcDNA3.1-C-HA vector using *NheI* and *XbaI*. We used the empty vector of pcDNA3.1-C-HA as a control. We cultured IMR-32 cells (ATCC, CCL-127) in Eagle's Minimum Essential Medium (ATCC, 30-2003) supplemented with 10% FBS (Corning, 35-079-CV). We nucleofected the IMR-32 cells with plasmids using a Lonza 4D-Nucleofector following the manufacturer's instructions (SF Cell Line 4D-Nucleofector X Kit, CA-168). Specifically, we overexpressed *FAM193B* by nucleofecting 4×10^5 IMR-32 cells with 1 μ g of *FAM193B* or control plasmid and 100 ng of pmaxGFP. For DNA damage analysis, we washed the cells three times with DPBS and fixed them with 4% paraformaldehyde 48 hours after nucleofection. We permeabilized and blocked the fixed cells with 0.4% Triton X-100, 1% BSA and 4% goat serum in DPBS for 1 hour at room temperature. We then stained the fixed cells with γ H2AX (Phospho-Histone H2A.X (Ser139)) by incubating cells with anti- γ H2AX (1:500, clone JBW301; Millipore, 05-636) overnight at 4 °C in DPBS with 4% goat serum, 0.4% Triton X-100 and 1% BSA. After primary antibody incubation, cells were incubated for 2 hours at room temperature with Alexa Fluor 568-conjugated goat anti-mouse IgG1 secondary antibody (1:1,000; Life Technologies, A21124) and DAPI (1:10,000; Life Technologies, D3571) in DPBS containing 0.02% Triton X-100 and 1% BSA. We performed imaging using a Leica TCS SP8 confocal system. For each biological replicate, we captured at least 15 images and used QuPath (version 0.5.1) to quantify the average number of nuclear γ H2AX foci in GFP⁺ cells. We compared the average numbers of γ H2AX foci between the *FAM193B* overexpression and control groups by unpaired two-tailed *t*-test. For RNA-seq, we sorted all GFP⁺ cells using a BD Influx cell sorter (BD Biosciences) and extracted total RNA using a NucleoSpin RNA Kit (Macherey-Nagel, 740955.50). We sent the RNA samples to Novogene for bulk RNA-seq on a NovaSeq X Plus platform. We processed RNA-seq data using the same analytical pipeline as the blood RNA-seq data (Supplementary Methods).

Differential gene expression analysis

We performed differential gene expression analysis on the single-nucleus RNA-seq data of the UKBBN samples to investigate the transcriptomic changes associated with AD ($n = 58$ patients with AD versus $n = 44$ NCs) or the *FAM193B*-CCG repeat ($n = 44$ NCs). Similar to eQTL analysis, we used the hurdle model from the MAST R package⁶⁶ to perform regression analysis with adjustment for age, sex, postmortem duration, gene detection rate and the first three principal components. To analyze the transcriptomic changes in *FAM193B*-overexpressing IMR-32 cells ($n = 6$ biological replicates per group), we kept genes with an expression level of TPM > 1 in at least half of the samples and used DESeq2 to perform differential gene expression analysis⁷⁰, adjusting for cell culture batches. We applied log(fold change) shrinkage to adjust noisy estimates using the lfcShrink(type = 'apeglm') function in DESeq2. Genes that passed multiple test correction (that is, FDR < 0.05) were considered significant differentially expressed genes and were subjected to Gene Ontology enrichment analysis (Supplementary Methods).

Data visualization

We visualized read alignments using Integrative Genomics Viewer (version 2.14.0) and the VizAln tool from HipSTR (version 0.6.2)^{24,71}. We visualized genome-wide association results using the qqman R package⁷² and locus-wise association results using LocusZoom (version 1.4)⁷³. We generated SMR plots using the script provided by the authors of SMR⁶⁸. We generated all bar, box and scatter plots as well as ROC curves with GraphPad Prism (version 8.0.2). For bar plots, bars show the mean value, and error bars indicate standard error of the mean. For box plots, boxes show the 25th to 75th percentiles (median center); whiskers mark the 10th and 90th percentiles; and points represent all

individual values. We generated density plots and heatmap plots using the ggplot2 R package⁷⁴.

Reporting summary

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

Data availability

The sources of publicly available PacBio data are presented in Supplementary Table 1. The SV genotypes of publicly available samples and AD GWAS summary statistics are available on Zenodo (<https://doi.org/10.5281/zenodo.14851280>). The LRS data of newly sequenced samples are available on Genome Sequence Archive (GSA) for Human (<https://ngdc.cncb.ac.cn/gsa-human/>) under accession number HRA015540, and their SV genotypes are available on Genome Variation Map (GVM; <https://ngdc.cncb.ac.cn/gvm/>) under accession number GVM001257. The repeat genotypes are available on GVM under accession number GVM001283; the RNA-seq data of IMR-32 cells are available on GSA for Human under accession number HRA015566. The UK Biobank data are available at <https://www.ukbiobank.ac.uk>. The GTEx data are available on the GTEx Portal (<https://gtexportal.org/home/>). The snATAC-seq data are available on the Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE147672. The ChIP-seq data are available on the ENCODE database (<https://www.encodeproject.org/>) under accession numbers ENCFF061CAX, ENCFF641AVN and ENCSR290SSQ. The RNA-seq data of the MAZ-knockdown cell line are available on the GEO under accession number GSE215080.

Regarding the other personally identifiable genome-wide and transcriptome-wide data, the consent form signed by each participant states that these data will remain private under the supervision of the hospital and research team. Therefore, these data will be made available and shared only in the context of a formal collaboration. Applications for data sharing and project collaboration will be processed and reviewed by a review panel hosted at the Hong Kong University of Science and Technology. Researchers may contact sklneurosci@ust.hk for further details on project collaboration and the sharing of data from this study. All requests will receive a response within 3 weeks. Source data are provided with this paper.

Code availability

The code and source data supporting this study are available on GitHub (https://github.com/Han-Cao/AD_Structural_Variation) and Zenodo (<https://doi.org/10.5281/zenodo.14851280>). The SV calling workflow is available at GitHub (<https://github.com/Han-Cao/HarmoniSV>). The other software are described in the Methods and Supplementary Methods sections along with citations.

References

- Nasreddine, Z. S. et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **53**, 695–699 (2005).
- American Psychological Association. *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition: DSM-5* (American Psychiatric Publishing, 2013).
- Pacific Biosciences. PacBio Secondary Analysis Tools on Bioconda. *GitHub* <https://github.com/PacificBiosciences/pbbioconda> (2026).
- Fukasawa, Y., Ermini, L., Wang, H., Carty, K. & Cheung, M. S. LongQC: a quality control tool for third generation sequencing long read data. *G3 (Bethesda)* **10**, 1193–1196 (2020).
- Kirsche, M. et al. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat. Methods* **20**, 408–417 <https://doi.org/10.1038/s41592-022-01753-3> (2023).
- Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

47. Zook, J. M. et al. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
48. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).
49. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
50. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
51. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
52. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://doi.org/10.48550/arXiv.1303.3997> (2013).
53. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
54. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
55. Privé, F., Luu, K., Blum, M. G. B., McGrath, J. J. & Vilhjálmsson, B. J. Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* **36**, 4449–4457 (2020).
56. Mousavi, N. et al. TRTools: a toolkit for genome-wide analysis of tandem repeats. *Bioinformatics* **37**, 731–733 (2021).
57. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
58. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
59. Canty, A. & Ripley, B. D. boot: Bootstrap Functions. R version 1.3-32 <https://doi.org/10.32614/CRAN.package.boot> (2022)
60. Thulin, M. boot.pval: Bootstrap p-Values. R version 0.7.0 <https://doi.org/10.32614/CRAN.package.boot.pval> (2025)
61. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
62. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
63. Inoue, E. nricens: NRI for Risk Prediction Models with Time to Event and Binary Response Data. R version 1.6 <https://doi.org/10.32614/CRAN.package.nricens> (2018)
64. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
65. de Klein, N. et al. Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nat. Genet.* **55**, 377–388 (2023).
66. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
67. Jang, B. et al. A meta-analysis of single-nucleus expression quantitative trait loci linking genetic risk to brain disorders. *Nat. Genet.* **58**, 737–747 (2026).
68. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
69. Yusa, K., Zhou, L., Li, M. A., Bradley, A. & Craig, N. L. A hyperactive piggyBac transposase for mammalian applications. *Proc. Natl Acad. Sci. USA* **108**, 1531–1536 (2011).
70. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
71. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
72. Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw* **3**, 731 (2018).
73. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
74. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).

Acknowledgements

We thank E. Chan (Chinese University of Hong Kong) for his comments and suggestions concerning the experimental validation of transcription factor binding. We thank R. M. N. Lo and B. W. Y. Wong for coordinating clinical data collection. We also thank L.-Y. Wong, K. C. Lok and W. S. Kwong for their technical assistance. This research was conducted using the UK Biobank Resource under application number 95743.

Author contributions

H.C., X.Z., A.K.Y.F. and N.Y.I. conceived of the study. F.C.F.I., Y.C. and T.C.Y.K. organized patient recruitment and sample collection. Y.F., Y.D. and S.H. performed the experiments. H.C. implemented the analytical tools. H.C. and S.C.L. set up the data processing pipelines. H.C., Y.F., X.Z., Y.D., K.Y.M., T.H.C., J.H., A.K.Y.F. and N.Y.I. analyzed the data. H.C., X.Z., A.K.Y.F. and N.Y.I. wrote the paper, with input from all authors.

Funding

This study was supported, in part, by the Research Grants Council (RGC) of Hong Kong (the Collaborative Research Fund (C6027-19GF), the Theme-Based Research Scheme (T13-605/18W) and the General Research Fund (HKUST16103122, HKUST16102824 and HKUST16104624)); the Areas of Excellence Scheme of the University Grants Committee (AoE/M-604/16); the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government (ITCPD/17-9, ITS/207/18FP and MRP/042/18X); the National Natural Science Foundation of China (NSFC)/RGC Joint Research Scheme (32061160472 and N_HKUST605/20); the Guangdong Provincial Fund for Basic and Applied Basic Research (2019B1515130004); and the Chow Tai Fook Charity Foundation (CTFCF18SC01). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests

J.H. has served as a consultant for Eli Lilly and Eisai. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41593-026-02481-z>.

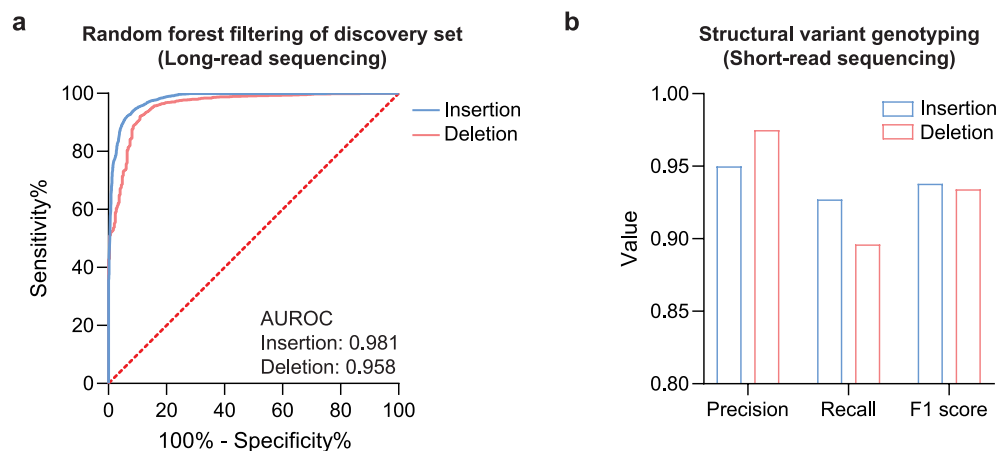
Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41593-026-02481-z>.

Correspondence and requests for materials should be addressed to Amy K. Y. Fu or Nancy Y. Ip.

Peer review information *Nature Neuroscience* thanks Julie Williams and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

All authors have read and approved the submission of this manuscript.

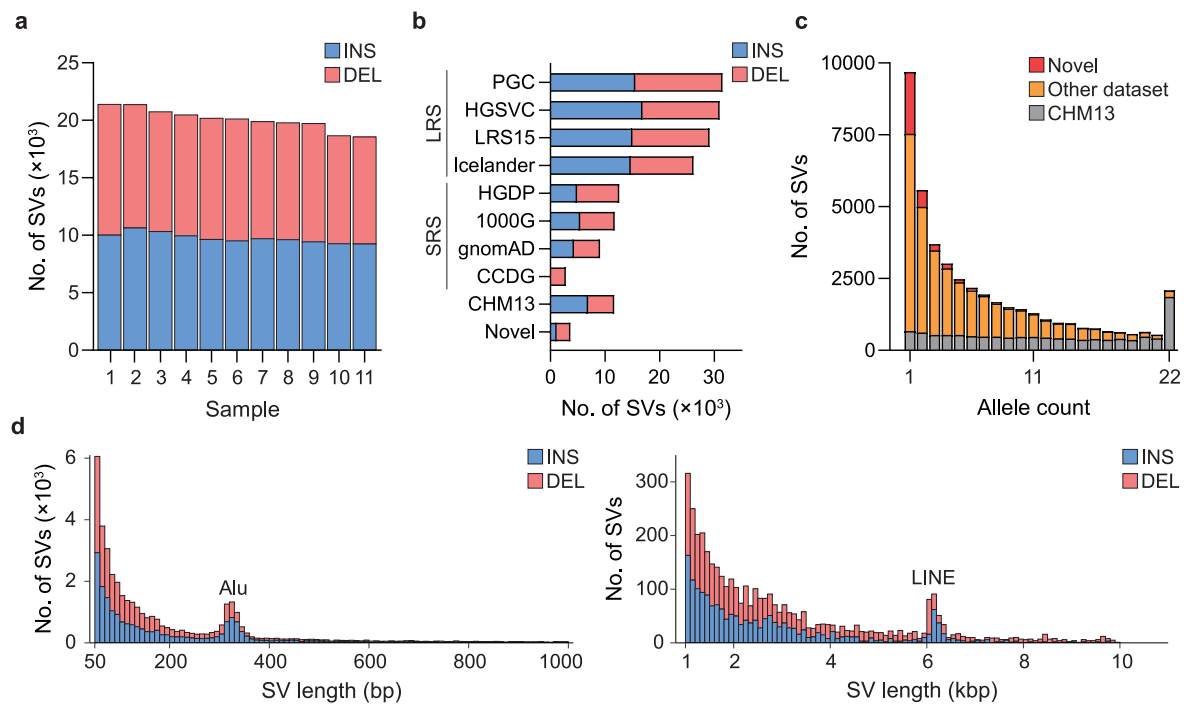


Extended Data Fig. 1 | Benchmark of structural variant calling and genotyping.

(a) Receiver operating characteristic (ROC) curve showing the performance of the random forest model for filtering false-positive structural variants (SVs)

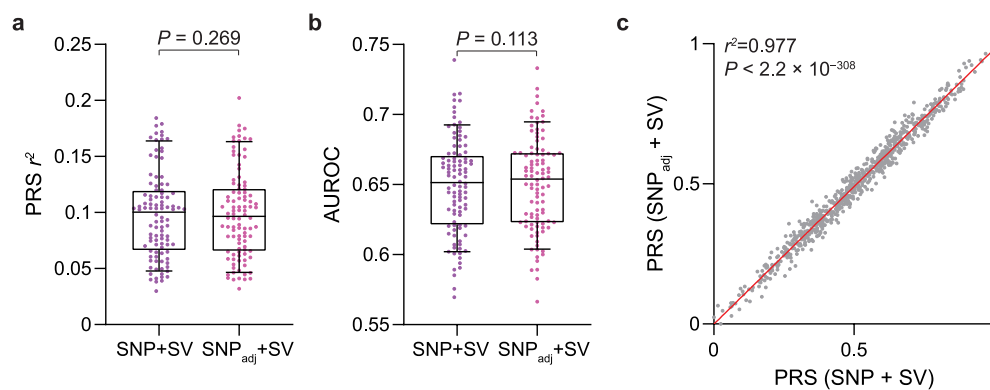
discovered from long-read sequencing data. AUROC, area under the ROC curve.

(b) Validation of the short-read sequencing-based SV genotyping results using PacBio HiFi sequencing data of one randomly selected sample.



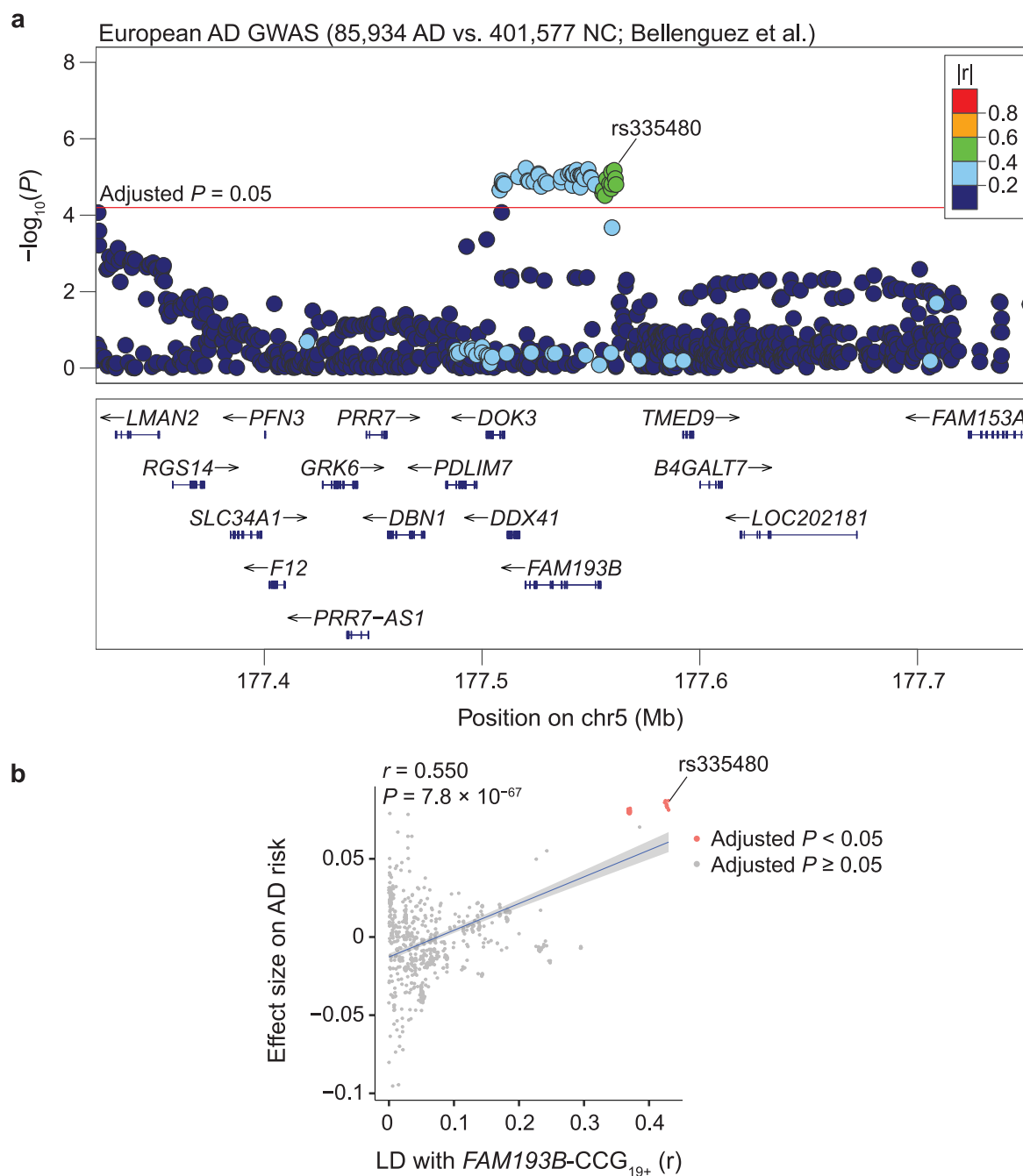
Extended Data Fig. 2 | Characterization of structural variants identified in the study cohort. (a) Number of structural variants (SVs) per genome. **(b)** Overlap between SVs identified in this study and other SV datasets. **(c)** Distribution of SV frequency. Known SVs present in public SV datasets and the CHM13 reference

genome are shown in yellow and grey, respectively. Novel SVs identified in this study are shown in red. **(d)** Distribution of SV length. DEL, deletion; INS, insertion; LINE, long interspersed nuclear element; LRS, long-read sequencing; SRS, short-read sequencing.



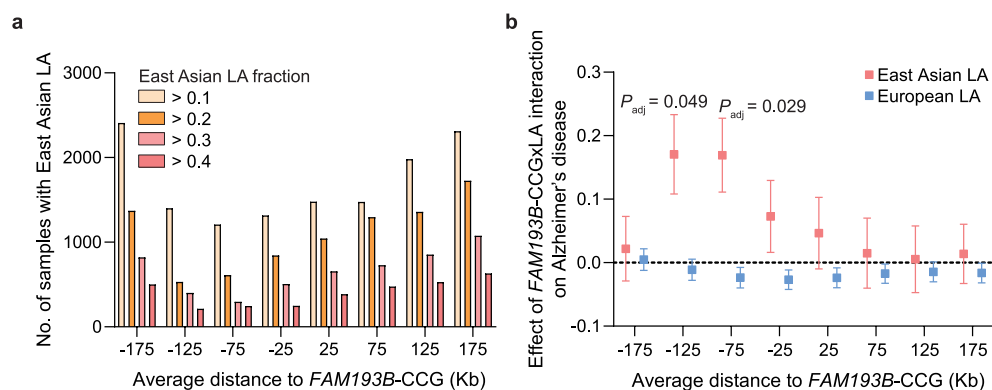
Extended Data Fig. 3 | Conditional polygenic risk score analysis for Alzheimer's disease. (a, b) Performance of SNP + SV and SNP_{adj} + SV polygenic risk score (PRS) models as measured by Nagelkerke's pseudo r^2 (a) and AUROC (b). Both models were constructed using SNPs and SVs, while the effects of SNPs in SNP_{adj} + SV models were further adjusted for nearby lead SVs. Ten times repeated 10-fold cross-validation. Paired two-tailed Wilcoxon signed-rank test.

Boxplots show medians, interquartile ranges, 10th and 90th percentiles, and all individual values. (c) Correlation between the PRSs calculated from SNP + SV and SNP_{adj} + SV models. r^2 , Pearson's correlation coefficient. AUROC, area under the receiver operating characteristic curve; SNP, single-nucleotide polymorphism; SV, structural variant.



Extended Data Fig. 4 | Linkage disequilibrium between *FAM193B*-CCG₁₉₊ and nearby genetic variants in the European population. (a) LocusZoom plot showing the associations between Alzheimer's disease (AD) and genetic variants in the *FAM193B* locus. Logistic regression. Filled colors denote the linkage disequilibrium (LD; as measured by $|r|$) between *FAM193B*-CCG₁₉₊ and other variants. The red line represents locus-wise significance threshold

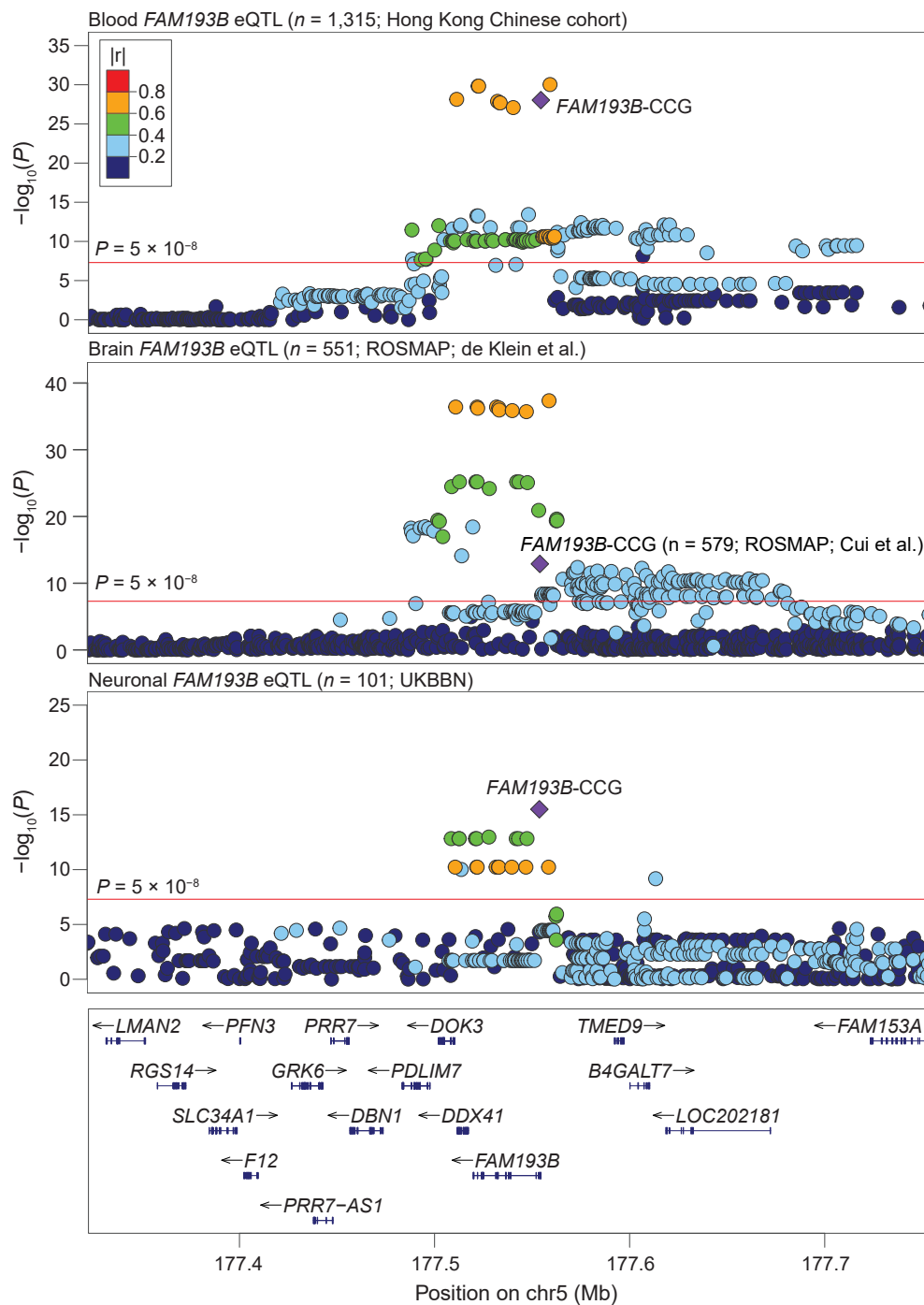
($P = 0.05 / 831 = 6.0 \times 10^{-5}$). The summary statistics for AD associations were obtained from Bellenguez et al. (doi:10.1038/s41588-022-01024-z). *FAM193B*-CCG₁₉₊, *FAM193B*-CCG alleles with ≥ 19 CCG repeats. r , Pearson's correlation coefficient. **(b)** Correlation between the AD risk effects of genetic variants in the *FAM193B* locus and their LD (as measured by r) with *FAM193B*-CCG₁₉₊. Two-tailed correlation test.



Extended Data Fig. 5 | Local ancestry analysis in the *FAM193B* locus.

(a) Number of samples with East Asian local ancestry (LA). The local ancestry of the 200-kb flanking region of the *FAM193B*-CCG repeat was estimated per 50-kb window for 2,728 clinically diagnosed Alzheimer's disease (AD) cases and 23,781 controls from the UK Biobank. Colors denote the minimum East Asian LA

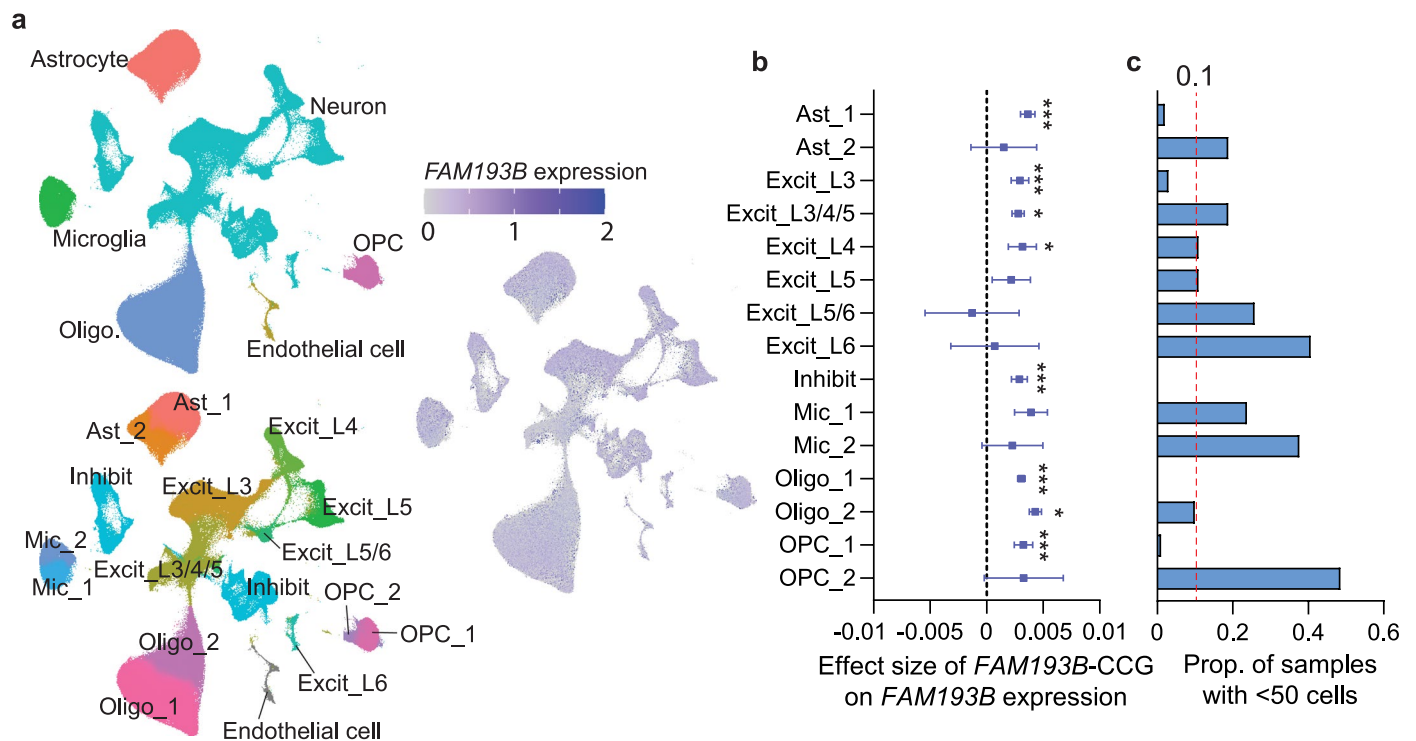
fraction. **(b)** Effects of the interaction term between the *FAM193B*-CCG repeat and LA on AD. $n = 2,728$ patients with AD and 23,781 controls. Whole-genome logistic regression. Data are presented as effect size $\pm$ standard error. P_{adj} , Bonferroni adjusted P values.



Extended Data Fig. 6 | Expression quantitative trait loci mapping of *FAM193B*.

LocusZoom plots showing the expression quantitative trait loci (eQTL) of *FAM193B* in blood (top), brain (middle), and neurons (bottom). Blood eQTL analysis was performed in the Hong Kong Chinese cohort. Brain eQTL summary statistics were obtained from Cui et al. (doi:10.1038/s41588-024-02057-2) and de Klein et al. (doi:10.1038/s41588-023-01300-6) for the *FAM193B*-CCG repeat and other genetic variants, respectively. Neuronal eQTL analysis was conducted

using the single-nucleus RNA sequencing data of samples obtained from the UK Brain Bank Network (UKBBN). Linear regression under an additive model for the top and middle panels, hurdle model regression under an additive model for the bottom panel. Filled colors denote the linkage disequilibrium (LD; as measured by $|r|$) between the *FAM193B*-CCG repeat and other variants. Red lines denote the genome-wide significance threshold.



Extended Data Fig. 7 | Regulatory effects of the FAM193B-CCG repeat on FAM193B expression in brain cells. (a) Expression of FAM193B in different brain cell subpopulations. Ast, astrocyte; Excit, excitatory neuron; Inhibit, inhibitory neuron; Mic, microglia; Oligo, oligodendrocyte; OPC, oligodendrocyte progenitor cell. (b) Associations between the FAM193B-CCG repeat and FAM193B

expression in brain cell subpopulations. $n = 101$. * Bonferroni adjusted $P < 0.05$, *** Bonferroni adjusted $P < 1 \times 10^{-4}$, hurdle model regression under an additive model. Data are presented as effect size $\pm$ standard error. (c) Proportion of samples with fewer than 50 cells for each brain cell subpopulation. Prop., proportion. $n = 101$.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	For the Hong Kong Chinese AD cohort, we recruited 2,333 participants from the Prince of Wales Hospital, Haven of Hope Christian Service, and Community CareAge Foundation. We also obtained genetic data and clinical records from a mainland Chinese AD cohort that has been previously described in Zhou et al. 2018 . This cohort includes 1,021 patients with AD and 867 NCs that are unrelated with each other. We obtained the genetic data and phenotype records of UK Biobank under application ID 95743.
Data analysis	The code and raw data to reproduce the results in this study is available on Zenodo (https://doi.org/10.5281/zenodo.14851280). We developed a workflow (https://github.com/Han-Cao/HarmoniSV) to performing SV calling by using the combination of 2 long-read aligners minimap2 (v2.28-r1209), NGMLR (0.2.7) and 3 SV callers Sniffles2 (2.2), cuteSV (2.0.3), and SVIM (2.0.0) We evaluated SV calls by using Truvari (v4.1.0) We genotyped FAM193B-CCG tandem repeat by using HipSTR (0.6.2), GangSTR (2.5.0), ExpansionHunter (v5.0.0), MergeSTR (6.1.0), and EnsembleTR (1.0.0) We used BCFtools (v1.11) to process the raw genetic data in Variant Call Format (VCF). We conducted genetic quality control analysis by using PLINK (v1.9, v2.0) and bigsnpr (v1.12.2) Genome-wide association analysis and polygenic risk score analysis were conducted by using PLINK (v1.9, v2.0) and REGENIE (v4.1.gz). We estimated the local ancestry in the FAM193B locus using SparsePainter (v1.3.2). We performed meta-analysis by using METASOFT (v2.0.0). We used Seurat (v4) to process all single-nucleus RNA sequencing data and used MAST (1.32.0) to perform regression analysis. We processed bulk RNA sequencing data using STAR (v2.7.10b), RSEM (v1.3.3) and performed differential gene expression analysis using DESeq2 (v1.42.1). We quantified gray matter volumes by using AccuBrain IV1.2 (BrainNow Medical Technology).

We quantified DNA probe intensity of EMSA using ImageJ (version 1.53q).
We quantified nuclear γ -H2AX foci using QuPath (v0.5.1).
All other statistical analysis were conducted by using functions in R (4.3.2) or Prism (8.0.2).

We visualized read alignments using Integrative Genomics Viewer (version 2.14.0) and the VizAln tool from HipSTR (0.6.2). We visualized genome-wide association results using the qqman R package and locus-wise association results using LocusZoom (v1.4). We generated SMR plots using the script provided by the authors of SMR. We generated all bar, box, and scatter plots as well as ROC curves with GraphPad Prism (version 8.0.2). We generated density plots and heatmaps using the ggplot2 R package.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The sources of publicly available PacBio data are presented in Supplementary Table 1. The SV genotypes of publicly available samples and AD GWAS summary statistics are available on Zenodo (<https://doi.org/10.5281/zenodo.14851280>). The LRS data of newly sequenced samples are available on Genome Sequence Archive (GSA) for Human (<https://ngdc.cncb.ac.cn/gsa-human/>) under accession number HRA015540, while their SV genotypes are available on Genome Variation Map (GVM; <https://ngdc.cncb.ac.cn/gvm/>) under accession number GVM001257. The repeat genotypes are available on GVM under accession number GVM001283; the RNA-seq data of IMR-32 cells are available on GSA for Human under accession number HRA015566. The UK Biobank data are available at <https://www.ukbiobank.ac.uk>. The GTEx data are available on the GTEx Portal (<https://gtexportal.org/home/>). The single-nucleus ATAC-seq data are available on the Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE147672. The ChIP-seq data are available on the ENCODE database (<https://www.encodeproject.org/>) under accession numbers ENCF061CAX, ENCF641AVN, and ENCSR290SSQ. The RNA-seq data of the MAZ-knockdown cell line are available on GEO under accession number GSE215080.

Regarding the other personally identifiable genome-wide and transcriptome-wide data, the consent form signed by each participant states that these data will remain private under the supervision of the hospital and research team. Therefore, these data will only be made available and shared in the context of a formal collaboration. Applications for data sharing and project collaboration will be processed and reviewed by a review panel hosted at the Hong Kong University of Science and Technology. Researchers may contact skneurosci@ust.hk for further details on project collaboration and the sharing of data from this study.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Sex were self-reported and further confirmed by genetic data. We adjust sex as a covariates for all association analysis. We did not perform any analysis stratified by sex.

Reporting on race, ethnicity, or other socially relevant groupings

We performed genetic analysis using genetic data from the Chinese and European populations. The ethnicity of participants were determined by genotypes.

Population characteristics

We have described the age, sex, clinical diagnosis of the study cohort in Supplementary Table 6.

Recruitment

For the Hong Kong Chinese cohort, we recruited 2,333 participants from the Specialist Outpatient Clinics of the Department of Medicine and Therapeutics at the Prince of Wales Hospital, Haven of Hope Christian Service, and Community CareAge Foundation. All participants underwent the Montreal Cognitive Assessment (MoCA), and patients with AD were diagnosed according to the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5). Participants were offered a HK\$250 supermarket voucher as a compensation token to support cohort recruitment and retention.

Ethics oversight

This study was approved by the Clinical Research & Ethics Committees of the Joint Chinese University of Hong Kong–New Territories East Cluster for the Prince of Wales Hospital as well as the Institutional Review Board, Hospital Authority (CREC ref. no. 2015.461, 2021.651), and the Human Participants Research Panel of the Hong Kong University of Science and Technology (CRP#180, HREP-2023-0179). All participants provided written informed consent for both study enrolment and sample collection.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences

☐ Behavioural & social sciences

☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size of different cohorts were described in Methods and Supplementary Table 6. We listed the sample size of individual analysis in Figure legends and Supplementary Tables. We used all samples with available data for analysis.
Data exclusions	Related participants are excluded from the analysis. Participants that are not of Chinese or European ancestry were also excluded. Participants whose genotypes fail quality control were excluded from specific analysis (see Methods and Supplementary Methods for details).
Replication	We used data from multiple independent cohort to replicate our findings.
Randomization	No Randomization was introduced in the analysis.
Blinding	The investigators who conducted the sample preparation and experiments were blinded for the data category

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems	Methods
n/a	n/a
☐ ☒ Antibodies	☒ ☐ ChIP-seq
☐ ☒ Eukaryotic cell lines	☒ ☐ Flow cytometry
☒ ☐ Palaeontology and archaeology	☐ ☒ MRI-based neuroimaging
☒ ☐ Animals and other organisms	
☒ ☐ Clinical data	
☒ ☐ Dual use research of concern	
☒ ☐ Plants	

Antibodies

Antibodies used	Mouse anti-Phospho-Histone H2A.X, Ser139 (05-636, Millipore, 1:500, JBW301, monoclonal); Goat anti-Mouse IgG1 Cross-Adsorbed Secondary Antibody, Alexa Fluor 568 (A21124, Life Technologies, 1:1000).
Validation	Both antibodies have been used in previously published studies on the human cell lines. The Mouse anti-Phospho-Histone H2A.X, Ser139 antibody has been validated by Yamashita et al., Cancer Sci. 2020 Jul;111(7):2431-2439; Sone et al., Nat Commun. 2014 Dec 9;5:5691; Kleiner et al., Nat Chem Biol. 2015 Oct;11(10):807-14. The Goat anti-Mouse IgG1 Cross-Adsorbed Secondary Antibody, Alexa Fluor 568, has been validated by Nahon et al., Stem Cell Research, 2023 Sep 1;71:103180, Lee et al., Nat Protoc. 2022 May;17(5):1266–305.

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	IMR-32 and HEK293T were purchased from American Type Culture Collection (ATCC)
Authentication	The IMR-32 was purchased from ATCC with catalog #CCL-127 and authenticated by the vendor. The HEK293T was purchased from ATCC with catalog #CRL-3216 and authenticated by the vendor.
Mycoplasma contamination	The IMR-32 and HEK293T were tested and confirmed to be negative
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used

Plants

Seed stocks	Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.
Novel plant genotypes	Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.
Authentication	Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.

Magnetic resonance imaging

Experimental design

Design type	Cross-sectional morphometry analysis
Design specifications	T1-weighted structural MRI data was used to examine the association between genetic variants and gray matter volume
Behavioral performance measures	None

Acquisition

Imaging type(s)	T1-weighted structural MRI
Field strength	3T
Sequence & imaging parameters	3D FFE (gradient echo) pulse sequence was used. The images were acquired from coronal view, with slice thickness 5mm, TE/TR/flip angle = 3ms/7ms/8. The acquired image matrix size was 240 x 25 x 240.
Area of acquisition	Whole brain
Diffusion MRI	☐ Used ☒ Not used

Preprocessing

Preprocessing software	The MRI data were preprocessed by AccuBrain IV1.2 (BrainNow Medical Technology Ltd), a brain quantification tool that performs brain structure and tissue segmentation and quantification in a fully automatic mode. Given the T1-weighted MRI data, several brain structures (e.g., hippocampus, lateral ventricle, amygdala, etc.) and three major brain tissues (i.e., white matter, gray matter, and CSF) are segmented automatically based on prior anatomical knowledge specified by experienced radiologists. The anatomical information is automatically transformed into the individual brain.
Normalization	For details please refer to Abrigo et al., Acta Radiologica 2018
Normalization template	For details please refer to Abrigo et al., Acta Radiologica 2018
Noise and artifact removal	For details please refer to Abrigo et al., Acta Radiologica 2018
Volume censoring	Not applicable

Statistical modeling & inference

Model type and settings	We used fix effect linear regression model for data analysis. We have listed the adjusted covariates in the Methods section and Supplementary Table 21
Effect(s) tested	We examined the association between genetic variants and gray matter volumes.
Specify type of analysis:	☒ Whole brain ☐ ROI-based ☐ Both
Statistic type for inference	Not applicable
(See Eklund et al. 2016)	
Correction	We only performed one association test for MRI data, so multiple test correction is not applicable here.

Models & analysis

n/a	Involvement in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☒	☐ Multivariate modeling or predictive analysis