Long-read sequencing reveals diverse haplotypes and common structural variants in Alzheimer's Disease GWAS loci

Niccolò Tesi[1,2,*,#], Alex Salazar[1,*,#], Gerard Bouland[1,*], Daniel Alvarez[1], Yaran Zhang[1], Lydian Knoop[1], Natasja M. van Schoor[3,4], Martijn Huisman[3,4], Sanduni Wijesekera[1], Jana Krizova[1], Betty M. Tijms[5,6], Everard G.B. Vijverberg[5,6], ADGC, Bonn, CHARGE, EADB, EADI, FinnGen, GERAD, GR@ACE/DEGESCO, PGC-ALZ, Marc Hulsman[1], Sven van der Lee[1,5,6], Marcel Reinders[2], and Henne Holstege[1,2,6,#]

1. Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands
2. Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands
3. Department of Epidemiology and Data Sciences, Amsterdam UMC location Vrije Universiteit Amsterdam, The Netherlands
4. Amsterdam Public Health Research Institute, Aging & Later Life, Amsterdam, The Netherlands
5. Alzheimer Center Amsterdam, Neurology, Vrije Universiteit Amsterdam, Amsterdam UMC location VUmc, Amsterdam, The Netherlands
6. Amsterdam Neuroscience, Neurodegeneration, Amsterdam, The Netherlands

* Contributed equally to this work

Corresponding authors: Niccolo Tesi (n.tesi@amsterdamumc.nl), Alex Salazar (a.n.salazar@amsterdamumc.nl), Henne Holstege (h.holstege@amsterdamumc.nl)

Target journals: Alzheimer's and Dementia – Genome Medicine – Molecular Neurodegeneration – Nature Neuroscience

Abstract (333 words)

Genome-wide association studies (GWAS) have identified over 100 Single Nucleotide Polymorphisms (SNPs) associated with Alzheimer's disease (AD) risk, however, most signals tag haplotypes rather than causal variants. This highlights the need to characterize haplotype-specific variation, including structural variants (SVs) and epigenetic modifications, as these may play a central role in shaping downstream disease mechanisms. We applied linkage disequilibrium (LD)-based clumping, followed by conditional analysis to identify significant and independent haplotypes associated with AD. Through long-read sequencing of 493 individuals, we systematically characterized the SV and DNA methylation landscape of these haplotypes. We integrated allele-specific differential methylation and chromatin organization to prioritize SVs likely contributing to disease mechanisms. Finally, we explored the feasibility of imputation approaches to predict SV size in 5,936 array-genotyped individuals.

Using AD-GWAS summary statistics for 98 GWAS loci we identified 280 independent and significant haplotypes. We then identified 2,000 unique SVs that were in LD (R^2>0.15) with 207/280 haplotypes. These SVs were predominantly composed of intronic transposable elements and tandem repeats, largely multi-allelic and overlapping regulatory regions. Based on differential methylation, genomic and chromatin co-localization, we prioritized 52 SVs as candidate contributors to disease mechanisms: 14 of these were in high LD with AD-haplotypes (R^2>0.8), 12 were in moderate LD (R^2>0.5), and 26 were in low LD (R^2>0.15). We identified intronic SVs in *TMEM106B, CYSTM1, IPMK, LMAN2, MINDY2*, as well as likely regulatory and exonic SVs in *APP, NDUFS2, TMEM184A*, *STRN4, CNN2, ADAM10,* and other loci. Fine mapping of the *PLEC/SHARPIN* locus revealed a novel haplotype with a tandem repeat expansion driving enhancer methylation and reduced *PLEC* expression in microglia. Finally, we imputed 83% of SVs with high accuracy (N=1,651, mean R^2=0.76), and association with AD status of imputed SVs yielded 112 significant associations (FDR<0.05).

AD risk loci are genetically complex, often comprising multiple haplotypes and linked SVs that could contribute to disease mechanisms. Integrating long-read sequencing, epigenetic data, and imputation strategies provides a more nuanced view of AD genetic architecture and highlights SVs as potential drivers of disease risk.

Introduction

Alzheimer's Disease (AD) is the most common form of dementia and one of major socio-economic burdens on society and patients' families [1]. Genetic factors play a major role in AD, with heritability estimates ranging 60-80% based on twin studies [2]. Our understanding of genetic factors underlying the most common form of AD (ie, late-onset AD), has been largely driven by genome-wide association studies (GWAS) [3]. To date, over 100 common Single Nucleotide Polymorphisms (SNPs) have been identified through GWAS, each contributing modestly to AD risk [3, 4]. These SNPs typically tag haplotypes characterized by multiple SNPs in linkage disequilibrium (LD) that co-occur together. It is increasingly evident that multiple SNPs in the same locus – the associated haplotypes – can independently contribute to disease risk, refining risk stratification and genetic association signals. For example, several AD-associated genes, including *SORL1, TREM2, SLC24A4, PLCG2,* and *MME,* contain multiple common risk haplotypes [3, 4]. Recognizing haplotype-level signals may provide a more nuanced understanding of AD genetic architecture beyond single-marker associations. Moreover, most SNPs identified through GWAS are not the direct causal variants, but act as proxies for other variants in LD within the same haplotype, including not only SNPs, but also larger structural variants (SV).

Long-read sequencing technologies combine longer read length, high accuracy, and epigenetic evaluation, and have facilitated the analysis of SV and other complex regions of the genome [5, 6]. The major types of SVs in the human genome are transposable elements (TE), and tandem repeats (TR). TE make up more than 30% of the human genome [7], largely due to their ability to self-replicate and integrate into other genomic loci. TRs are similarly widespread in the human genome and are composed of one or more adjacently repeating DNA sequences [8]. TRs have been implicated in various neurological diseases, including Huntington's disease, fragile X syndrome, and several forms of ataxia [8–10]. As is often the case in disease-associated SVs, expansions or contractions disrupt nearby gene-function, triggering a dysfunctional cascade. For AD, previous studies using short-read sequencing or experimental approaches have identified a handful of SVs influencing AD-risk [11–13].

80 In addition to genetic variation, epigenetic mechanisms such as DNA methylation at CpG sites are key
81 regulators of gene expression [14, 15]. Differential methylation patterns have been reported in genes
82 associated with AD pathogenesis, AD severity, neuroimaging findings, and increasing age [15, 16]. SNPs
83 and SVs within or near CpG sites can alter the affinity of DNA methyltransferases or transcription factors,
84 modify chromatin organization and regulatory elements, leading to inter-individual differences in gene
85 regulation [17–20].
86
87 Altogether, characterizing haplotypes by SNPs, SVs and integrating methylation status will contribute to a
88 deepened understanding of *how* risk haplotypes might exert their effects. For this, our recent observations
89 in the locus surrounding *TMEM106B* gene serves as proof of principle. This locus has been identified in
90 GWASs for AD and other neurological disorders [3]. The haplotype tagged by the AD-SNP rs13237518
91 harbors a coding substitution (T185S), historically suspect to dysregulate *TMEM106B* expression [21].
92 However, functional studies assessing the transcriptional and protein-level effects of this variant remain
93 inconclusive [22]. Recently, we and others have identified that the risk-haplotype co-segregates with an
94 insertion of a 300bp AluYb8 retrotransposon (23). This TE alters the local epigenetic landscape, likely
95 contributing to (post-)transcriptional mechanisms that dysregulate *TMEM106B* expression [23]. Thus, the
96 SNP is merely a proxy for other genetic variants, including a large SV that may have a greater impact on
97 regulatory architecture.
98
99 A systematic investigation of SVs, methylation effects and their interaction with AD risk-haplotypes is
100 essential to understand genetic mechanisms underlying AD. Here, we link the most common AD-associated
101 risk haplotypes with SVs and differential methylation patterns. To achieve this, we leveraged published AD-
102 GWAS summary statistics, long-read sequencing data from 493 individuals, and genotyping data from
103 5,936 individuals spanning the entire spectrum of cognitive decline, including relatively young AD patients,
104 healthy older adults, and cognitively healthy centenarians who reached 100 years without AD.
105

Results

Selection and definition of AD-haplotypes

We extracted summary statistics for 120 variants in 98 genomic loci previously associated with AD from published GWAS (Table S1), along with all SNPs located within 250 kilobases (kbp) of each locus (Figure 1A-B). To reduce redundancy and facilitate downstream haplotype analysis, we first grouped variants into LD-based clumps (see Methods) and identified common haplotypes with suggestive association with AD (p<5e-5). This approach yielded 532 haplotypes in AD loci (Figure 1C-D, Table S2). The number of SNPs per haplotype ranged between 0 and 3869 (mean=115, median=47). In our dataset, the regions near *ABCA7* and *KCNN4* had the highest number of haplotypes (36 and 38, respectively), whereas loci such as *OTULIN, TMEM184A, JAZF1, DOC4A, TPCN1, GGA2, SSBP4,* and *SRC* consisted of only one. For each of the 532 haplotypes, we selected a representative SNP, the one with the most significant association with AD, based on the AD-GWAS summary statistics.

To identify conditionally independent haplotypes associated with AD, we applied conditional and joint association (COJO) analysis (see Methods), which adjusts SNP-level p-values by accounting for LD with neighboring variants. Because each SNP tags a haplotype, this step effectively identifies haplotypes that independently associate with AD. In total, 280 haplotypes remained significantly associated after COJO analysis (False Discovery Rate <1%) (Figure 1C, Table S3). Throughout this manuscript, we refer to this final set of haplotypes as *AD-haplotypes*. Out of the 98 loci, the region near *ATP8B4* was excluded due to low minor allele frequency (<1%), nineteen loci (19%) had only one AD-haplotype, and 78 loci (80%) contained at least two AD-haplotype (mean=3.7, median=3, IQR=3). The loci surrounding *KCNN4* and *PLCG2* had the largest number of AD-haplotypes (10), followed by *ABCA7* (9), and *ADAM10* (8).

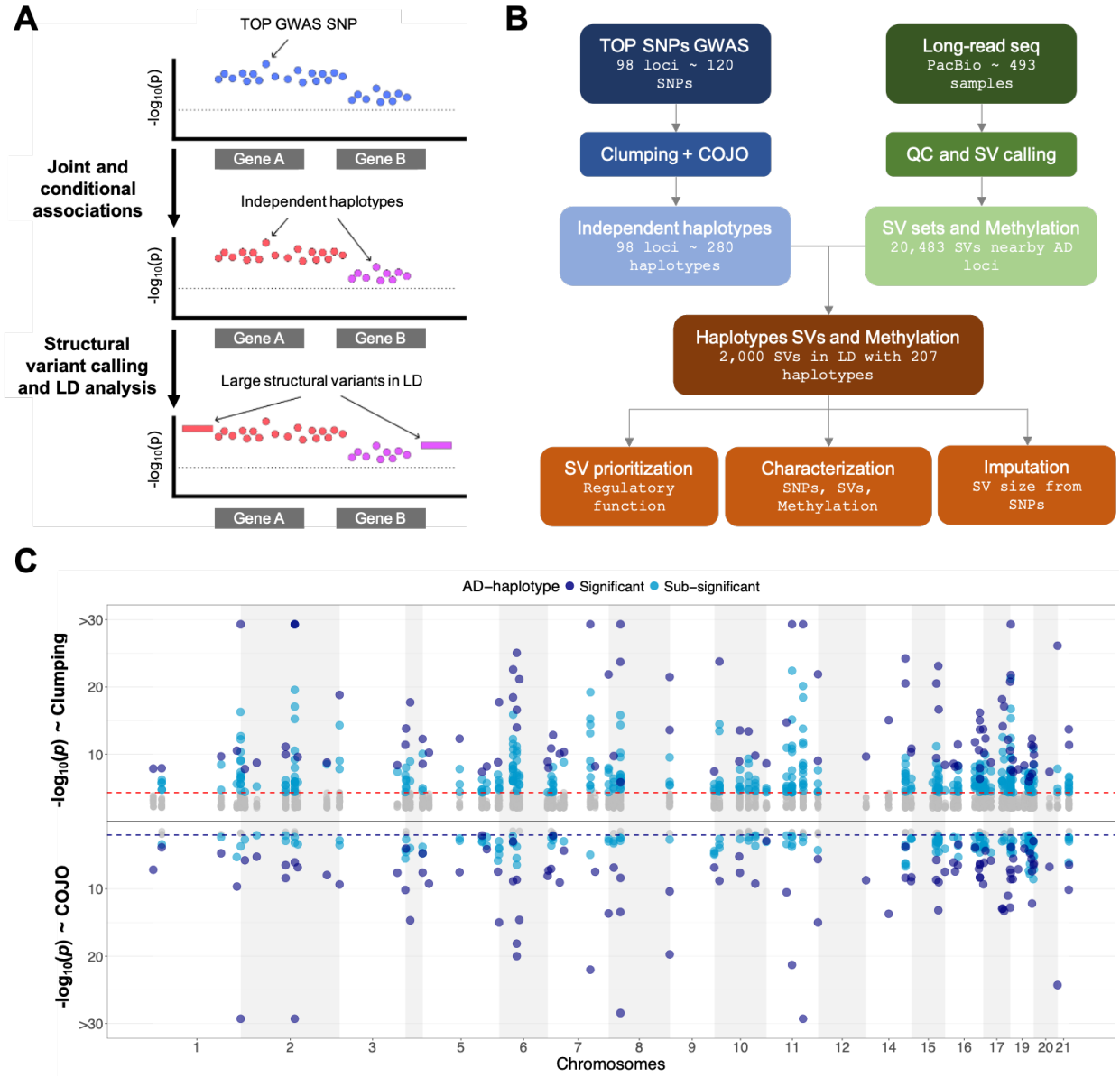


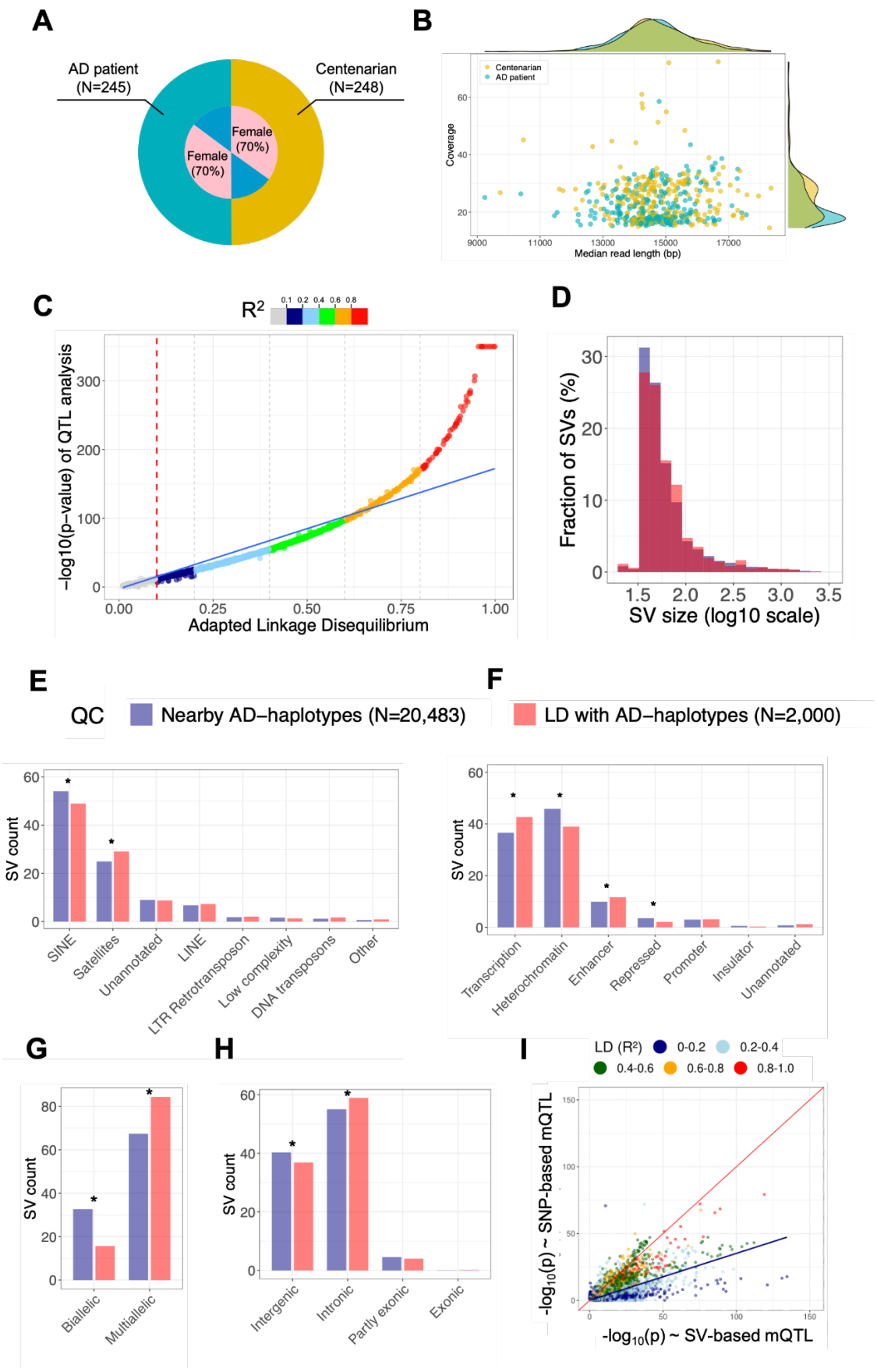
Figure 1: Analysis workflow and identification of AD-haplotypes. **A.** In GWAS analyses, the leading variant is often considered as the causal variant. However, other SNPs as well as Structural Variants (SV) within the same haplotype could explain the biological consequences underlying the association. Additionally, multiple haplotypes (each consisting of additional SNPs and SVs) may be present within each GWAS locus, which could also explain part of the association signal. **B.** Using GWAS summary statistics, through clumping and COJO analyses, we identified common haplotypes significantly associated with AD-risk. Using long-read sequencing data, we genotyped SVs and methylation signatures across 493 Dutch individuals. Then, we linked SVs to AD-haplotypes. Finally, we (*i*) prioritized SVs potentially contributing to disease mechanisms by integrating differential methylation signal, chromatin organization, and genomic context; (*ii*) characterized haplotypes in terms of their SNP, SV, and methylation content; and (*iii*) evaluated the

139 feasibility of imputation approaches to predict SV size from SNP genotypes. **C.** Miami plot showing haplotypes after
140 clumping and COJO procedures. Navy dots indicate known haplotypes associated with AD at genome-wide
141 significance. Light blue dots indicate the sub-significant (clumping) and independent (cojo) haplotypes. The red dashed
142 line indicates the threshold used during clumping (p<5e-5), while the blue dashed line indicates the threshold used
143 during COJO (False Discovery Rate p-value<1%).
144

145 Long-read sequencing of cognitively healthy centenarians and Alzheimer's 146 disease patients

147 We used PacBio technology for whole-genome sequencing of peripheral blood of 245 Alzheimer's disease
148 patients (AD, mean age 67.11, IQR=12, 70% females) and 248 centenarians (mean age 101.2, IQR=2,
149 70% females, Figure 2A). The median sequencing coverage, including both HiFi and non-HiFi (read quality
150 >0.85), was 25.7x across all individuals, with a median read length of 14,616 bp (Figure 2B). Although the
151 median read length did not differ between AD cases and centenarians (p=0.10, Wilcoxon rank-sum test),
152 the centenarian genomes had significantly higher coverage compared to AD genomes (AD: median 23.88x,
153 IQR=9.77; centenarian: median 27.76x, IQR=12.56; $p=1.07x10^{-4}$, Wilcoxon rank-sum test).
154
155 Combining structural variant calling and joint genotyping (see Methods), we identified 910,901 unique
156 structural variants (SVs) across all AD and centenarian genomes. After quality control (see Methods), we
157 identified 20,483 variable SVs located within 250 kbp of the 280 AD-haplotypes. The median size of this
158 set of SVs in the reference genome GRCh38 was 50 bp (mean=89.7 bp, Figure 2C).
159
160 The majority of the 20,483 SVs were intronic (55%) and intergenic (40%), while only 5% of SVs were exonic
161 (based on overlap with RefSeq gene track from UCSC) [24, 25]. Annotation of these SVs based on
162 RepeatMasker (obtained from UCSC Genome Browser) [24, 26] showed that 11,063 SVs (54%) mapped
163 to SINE-order elements, which are Transposable Elements (TE) that use an mRNA intermediate to replicate
164 and integrate elsewhere in the genome. The great majority of SINE corresponded to Alu-elements (99%),
165 reflecting recent evolution of the human genome. SINEs were followed by 5,105 Satellite elements (25%,

166 Figure 2E), including tandem repeats (TR) with various sizes of the repetitive motif (17% homopolymers,
167 35% dinucleotides, 8% trinucleotides, 24% tetranucleotides, and 16% with a motif >4 base pairs). LINE
168 elements (Long Interspersed Nuclear Elements), another major class of TE in humans, represented 7% of
169 SVs. Finally, 2% of SVs overlapped with LTR (Long Terminal Repeat) Retrotransposons, 2% overlapped
170 Low Complexity regions, 1% were DNA transposons, <1% overlapped other classes, and 9% of SVs did
171 not overlap with any annotation from RepeatMasker (Figure 2E). Annotation of the 20,483 variable SVs
172 based on chromatin marks [24, 27] (see Methods) showed that 9,387 SVs (46%) mapped to
173 heterochromatin regions, followed by regions associated with transcription (37%, 7,496), enhancers (10%,
174 2,016), and promoters (3%, 602) (Figure 2F). SVs are known to be highly variable, and they can be
175 characterized by multiple alleles. Joint genotyping of the 20,483 SVs across all samples (see Methods),
176 revealed that 33% were bi-allelic, and 67% were multi-allelic (Figure 2G), with more than 3 unique allele-
177 sequences across all samples. The median number of unique allele sequences across SVs was 5
178 (mean=7.0, min=2, max=349).
179
180 Overall, we curated and annotated a set of 20,483 variable SVs, each mapping in the vicinity of AD-
181 associated haplotypes.
182


A
AD patient (N=245)
Centenarian (N=248)
Female (70%)
Female (70%)
B
Centenarian
AD patient
Coverage
Median read length (bp)
C
R2
0.1 0.2 0.4 0.6 0.8
−log10(p−value) of QTL analysis
Adapted Linkage Disequilibrium
D
Fraction of SVs (%)
SV size (log10 scale)
E
QC
Nearby AD−haplotypes (N=20,483)
SV count
SINE
Satellites
Unannotated
LINE
LTR Retrotransposon
Low complexity
DNA transposons
Other
F
LD with AD−haplotypes (N=2,000)
SV count
Transcription
Heterochromatin
Enhancer
Repressed
Promoter
Insulator
Unannotated
G
SV count
Biallelic
Multiallelic
H
SV count
Intergenic
Intronic
Partly exonic
Exonic
I
LD (R2) 0-0.2 0.2-0.4 0.4-0.6 0.6-0.8 0.8-1.0
-log10(p) ~ SNP-based mQTL
-log10(p) ~ SV-based mQTL

183

184 Figure 2: Long-read sequencing. **A.** Pie chart showing the fraction of Cognitively Healthy Centenarians (N=248) and
185 AD patients (N=245), and the relative sex content. AD patients were from the Amsterdam Dementia Cohort, while
186 Cognitively Healthy Centenarians were from the 100-plus Study cohort [28, 29]. **B.** Scatterplot and marginal densities
187 of the sequencing coverage and read length of all 493 individuals. For all individuals, we used both HiFi and non-HiFi
188 (read quality>0.85) data. **C.** SVs in LD with AD haplotypes. Each dot represents a SNP-SV pair: the x-axis shows the
189 LD R^2 value calculated by scaling SV size in [0, 2] interval. The y-axis shows the -log10(p-value) of a QTL approach,
190 where the SV size was considered the quantitative trait and the SNP was the predictor. The Spearman correlation
191 coefficient between -log10(p-value) and R^2 was 0.99. **D.** Histogram of the Structural Variant (SV) size of the 20,483
192 variable SVs nearby AD-haplotypes (blue), and the SVs in linkage disequilibrium (LD) with AD haplotypes (2,000, red).
193 For visibility purposes, SV allele size is shown in log10 scale. **E.** Distribution of the major SV types observed in the
194 20,483 SVs nearby AD-haplotypes (blue) and the subset of 2,000 SVs in LD with AD haplotypes (red). Annotation is
195 based on RepeatMasker: when multiple annotations overlapped the SV of interest, the annotation with the largest
196 overlap was prioritized. **F.** Distribution of the major chromatin marks overlapping the full set of 20,483 SVs (blue) as
197 well as the subset of 2,000 SVs in LD with AD haplotypes. Chromatin annotation refers to the chromHMM 15-states
198 marks. **G.** Fraction of biallelic (*i.e.* allele counts <=3) and multiallelic SVs in the full set of 20,483 SVs as well as the
199 subset of 2,000 SVs in LD with AD haplotypes. SVs were classified as biallelic or multiallelic based on the unique alleles
200 as observed in our cohort of 493 samples after the joint-genotyping procedure. **H.** Genic context of SVs in the full set
201 of 20,483 SVs (blue) as well as the subset of 2,000 SVs in LD with AD haplotypes. **I.** Comparison of methylation-QTL
202 (mQTL) strength between SNP-based mQTLs (-log10(p) shown on y-axis) and SV-based mQTL (-log10(p) shown on
203 x-axis). *****: significant differences between the full set and the subset of SVs in LD with AD-haplotyped (Fisher's exact
204 test).
205

206 Structural variations are in linkage with AD-associated GWAS SNPs

207 To identify SVs in LD with AD-haplotypes, we combined two approaches. First, we performed a quantitative
208 trait locus (QTL) analysis between AD-haplotypes and SV lengths, using SV lengths as the quantitative
209 trait, and the AD-haplotypes (i.e the 280 representative AD-SNPs) as predictors (see Methods). While this
210 approach can identify SNP-SV pairs that are strongly associated, it does not provide a typical LD value
211 (such as R^2). Therefore, we also calculated a typical LD score by scaling the SV lengths for each region in

212 the interval [0-2] (see Methods), effectively scaling the SV lengths to the typical dosage of imputed SNPs.
213 This allowed us to calculate typical LD values between SNPs and SVs.
214
215 In total, we calculated 45,514 QTLs between 20,483 variable SVs and the 280 AD-haplotypes. Of these,
216 4,361 QTLs were significant (p<5e-8), and reflected associations between 242 AD-haplotypes (86% of the
217 280 AD-haplotypes), and 3,691 unique SVs (18% of all SVs nearby AD-haplotypes). For all AD-haplotypes
218 and SVs, we also calculated a typical LD value based on the scaled SV lengths. The resulting R^2 values
219 were highly correlated with the -log10(p) of the QTL (Spearman's correlation=0.99, p<2e-16), suggesting
220 that the two methods aligned (Figure 2C). We combined the two analyses and prioritized 2,125 SNP-SV
221 pairs (comprising 207 AD-haplotypes and 2,000 unique SVs) for which the QTL p-value was <5e-8 and the
222 R^2 value was >0.15 (Figure 2C, Table S4). Compared to the full set of 20,483 SVs, the subset of SVs in
223 AD-associated loci was significantly enriched with intronic SVs (OR=1.18 [1.07-1.29], p=8.9e-4) and
224 depleted with intergenic SVs (OR=0.86 [0.78-0.95], p=2.8e-3), while the proportion of exonic SVs did not
225 significantly differ (p=0.14) (Figure 2H). SVs in LD with AD-haplotypes were enriched for satellite SVs
226 (based on RepeatMasker annotation, OR=1.23 [1.12-1.36], p=3.4e-5, Fisher's exact test) and depleted with
227 SINE elements (OR=0.82 [0.74-0.89], p=9.2e-6) (Figure 2E) [26]. In terms of chromatin marks, this set of
228 SVs was enriched for enhancer marks (OR=1.22 [1.04-1.40], p=9.1e-03, Fisher's exact test) and
229 transcription-related marks (OR=1.30 [1.18-1.43], p=4.5e-8, Fisher's exact test), and depleted with
230 heterochromatin marks (OR=0.76 [0.69-0.84], p=9.7e-9, Fisher's exact test), and repressed marks
231 (OR=0.59 [0.42-0.82], p=6.9e-4, Fisher's exact test) (Figure 2F) [27]. The median size of this refined set of
232 SVs in GRCh38 was 52 bp (IQR=34, mean=89.3 bp, Figure 2D). SVs in LD with AD haplotypes were also
233 more likely to be multiallelic compared to the complete set of SVs (OR=2.61 [2.31-2.97], p=1.0e-61, Fisher's
234 exact test, median number of unique allele sequences was 7, IQR=6, mean=9.7, Figure 2G).
235
236 We evaluated how SVs may perturb the local methylation landscape (see Methods). For 18% of SVs (352
237 SVs), we observed significant allele-specific differential methylation after correcting for multiple testing
238 (FDR<0.05, Table S5), indicating that methylation frequency differed significantly between alleles of the
239 same SV. We then performed methylation-QTL (mQTL) analyses for both SVs and AD-haplotypes (see

240 Methods). When comparing SNPs and SVs in LD within the same haplotype, SV-based mQTL signals were
241 consistently stronger than those of SNPs (Figure 2I).
242
243 Altogether, we highlighted 2,000 SVs in LD with 207/280 (74%) of AD-associated haplotypes. Interestingly,
244 135/280 (48%) of the AD-haplotypes are expression- or splicing-QTLs (based on GTEx, see Methods) [30].
245 As SVs have known regulatory roles, SNP-QTLs may extend to t haplotype-specific gene-expression levels.
246 Indeed, we found that AD-haplotypes in LD with SVs were 2-fold more likely to be eQTLs based on the
247 GTEx database (p=0.01, OR=2.01 [1.12-3.65], Fisher's exact test).
248

249 Complex interplay of Structural variants and SNPs

250 For 46/207 haplotypes (22%, from 38 loci), we found at least one SV in high LD (R^2>0.8) with the haplotype;
251 for 72/207 haplotypes (35%, from 72 loci) the LD was moderate (R^2>0.5); and for 89 haplotypes (43%, from
252 89 loci), the LD was low (R^2>0.15). To identify SVs potentially contributing to molecular mechanisms, we
253 integrated allele-specific differential methylation, genomic context, and chromatin organization (see
254 Methods). SVs with a significant differential methylation, intronic or exonic localization, and overlapping
255 known regulatory regions such as enhancers, promoters, or transcriptionally active regions, were ranked
256 highest. This systematic approach enabled stratification of SVs into three tiers of regulatory relevance, with
257 SVs in *tier 1* (Figure 3A) representing the most compelling candidates, while *tier 2* and *tier 3* SVs showing
258 progressively lower levels of supporting evidence.
259

260 *SVs in high LD with AD-haplotypes*

261 Of the 46 AD-haplotypes-SV pairs in high LD, 14 SVs (30%) were prioritized as tier 1, 7 as tier 2, and 25
262 as tier 3 (Table 1). As expected, the known AluYb8 transposable element in *TMEM106B* was prioritized as
263 tier 1. The AD-associated variant rs6891922 (chr5:140,264,131, T>G, pAD=5.4e-07) is intronic of *PFDN1*,
264 has low pathogenicity score (CADD=2.7) [31], and acts as a strong eQTL and sQTL for the *CYSTM1* in
265 Brain and other tissues (based on GTEx) [30]. No coding variants lie on the same haplotype. We identified

266 a biallelic AluYb8 transposable element intronic of the gene, located in a transcriptionally active region and
267 in complete LD with rs6891922 (R^2=0.98). The risk allele of rs6891922 (G) co-occur with the presence of
268 the AluYb8, in which the expanded allele significantly increased local methylation compared to the shorter
269 (FDR<5e-150) (Figure 3B).
270
271 In the haplotype tagged by rs2790214 (chr10:58,253,643, C>T; pAD = 1.1e-09), the top SNP is an intronic
272 variant of *IPMK* (Figure 3C) with low pathogenicity score (CADD=3.2) [31]. Rs2790214 is a strong eQTL
273 for *IPMK* and *CISD1* genes in several tissues, based on GTEx [30]. We identified a structural variant in
274 complete LD with the top SNP (75-120 bp, chr10:58,268,914-58,269,009) located intronically of *IPMK* within
275 an active promoter (Figure 3C). The AD risk allele of rs2790214 (C) co-occur with larger SV size, which is
276 also associated with differential methylation (p<5e-150) supporting a role for the SV in interfering with the
277 local epigenetic landscape, and potentially *IPMK* expression.
278
279 The haplotype tagged by rs4131290 (chr5:177,353,681, A>C; pAD = 1.1e-07) is characterized by an
280 intergenic variant between *LMAN2* and *RGS14,* which is also a strong eQTL for *RGS14*, *MXD3*, and *LMAN2*
281 genes. This variant is in moderate LD with a synonymous coding variant in *NSD1* with low predicted
282 pathogenicity (rs11740250, R^2 = 0.57, CADD = 2.9) [31] and in high LD with a polymorphic (AC)n simple
283 repeat (60-120 bp; chr5:177,343,547-177,343,611, R^2=0.86), located intronically of *LMAN2* in a
284 transcriptionally active region. The AD risk allele (A) is in LD with shorter repeat alleles, which are
285 associated with lower methylation frequency (p<5e-150).
286
287 For 11 additional AD-haplotypes we observed similar structures, with the SV located in active promoters,
288 such as in the haplotypes tagged by rs12716979 (chr16:31,073,061-31,073,099 in *ZNF668*) and rs2289702
289 (chr15:78,943,234-78,943,289, AluS, in *CTSH*), or intronic within transcriptionally active regions, such as
290 in the haplotypes targeted by rs60073596 (SVA retrotransposon in *TOM1L2*), rs2154481 (AluSx3 in *APP*),
291 and rs7218117 (Unknown SV in *MINK1*).
292

SVs in moderate LD with AD-haplotypes

From the SVs in moderate LD (R^2>0.5) with 72 AD-haplotypes, 12 (16%) were classified as tier 1, 15 (21%) as tier 2, and the remaining 45 as tier 3 (Table 1). For example, in *TMEM184A* we identified an exonic biallelic SV (chr7:1,549,392-1,549,324, 23-57 bp) in moderate LD (R^2 = 0.56) with rs4257937 (chr7:1,552,361, G>C), an intronic variant with low pathogenicity (CADD=0.68) and a strong eQTL for *TMEM184A* in several tissues (based on GTEx) [30–32]. The SNP risk allele (C) is in LD with the expanded allele of the SV, which significantly increases methylation frequency (p<5e-150), suggesting a potential mechanistic effect on gene expression (Figure 3E). In *NDUFS2* and *STRN4* genes, the haplotypes tagged by variants rs4379692 (chr1:161,216,523, C>T, pAD = 3.5e-09) and rs146335962 (chr19:46,651,227, C>T, pAD = 6.1e-06) are both in moderate LD with exonic repeats: a (TG)n repeat ranging 60-78 bp in size in *NDUFS2* (R^2 = 0.59 with rs4379692); and a larger (CCCG)n repeat of size 96-103 in *STRN4* (R^2 = 0.65 with rs146335962), localized within an active promoter and with expanded alleles significantly increasing methylation frequency (p<5e-150). In *EIF4G3, MGAT5, SPON2, FAM193B, GPC2, PLEC, CDK10* genes, the SV in the respective haplotypes colocalized within enhancer or transcriptionally active regions, and reported significant allele-specific differential methylation, suggesting a potential perturbation of the epigenetic landscape and a potential effect on gene expression.

SVs in low LD with AD-haplotypes

We finally examined haplotypes where the leading AD-associated variant and SVs were in low LD (R^2<0.5). While these SVs are not consistently occurring with the leading risk variants, they may still contribute to disease mechanisms. In total, 29% (26/89) of these SVs were classified as tier 1, 33% as tier 2, and the remaining 42% as tier 3. In *INPP5D* gene, we identified a biallelic AluSz transposable element (chr2:233,199,117-233,199,559, 293-443 bp in size) in low LD with rs13007032 (R^2 = 0.28), located in a transcriptionally active region, with increasing AluSz size co-occurring with the protective allele, and associating with increased methylation (p<5e-150). In *RASGEF1C* gene, a 332-387 bp (CCCGCCG)n repeat was in low LD (R^2 = 0.22) with rs113706587 (chr5:180,201,150, G>A), which is also a known eQTL for the gene in Brain (based on GTEx). Expanded alleles of this CG-rich repeat increase local methylation

levels, contributing to the strong allele-specific differential methylation we observed ($p<5e-150$). Similarly, in *ABCA1* gene, the intronic rs2297400 was in LD with a large, biallelic L1ME4c transposable element (size ranging 55-1,015 bp) intronic of *ABCA1*, with the expanded allele associated with significant increased in methylation frequency. Finally, a common haplotype in *TREML2* (tagged by rs34346157) was in LD (R^2 = 0.46) with an exonic (TGTC)n repeat, with the rs34346157-G allele associated with increased AD-risk and with a larger TR size. Similar cases were also observed in *USP6NL, SLC24A4* (repeat of (AC)n of size 70-105 bp)*, ZNF652* (exonic biallelic transposable element), and *CNN2* (intronic repeat ranging 500-2.250 bp, Figure 3D).

Of note, in some cases, the same SV was in LD with multiple AD-haplotypes. For example, a biallelic SV intronic of *ADAM10* (chr15:58,699,523-58,699,595, 59-73 bp), which overlaps an enhancer and carries allele-specific differential methylation, was found in low LD with the haplotypes tagged by rs2657125 (R^2 = 0.22) and rs442495 (R^2 = 0.18), which are independent of each other (R^2 = 0.05). The (TT)n homopolymer repeat exonic of *CNN2* (chr19:1,037,926-1,037,995, 67-79 bp) overlaps a strong enhancer and is in LD with two haplotypes in low LD (R^2 = 0.12) in the region, tagged by rs75364577 (R^2 with SV was 0.32) and rs12151021 (R^2 with SV was 0.22). We observed similar cases in/near *PLEC, SLC24A4, PTK2B, SNX1, MS4A4A,* and 31 additional loci.

Taken together, in 25% of the 207 AD-haplotypes containing at least one SV, these are candidate contributors to disease mechanisms based on our prioritization method and could be altering the expression of nearby genes through regulatory mechanisms.

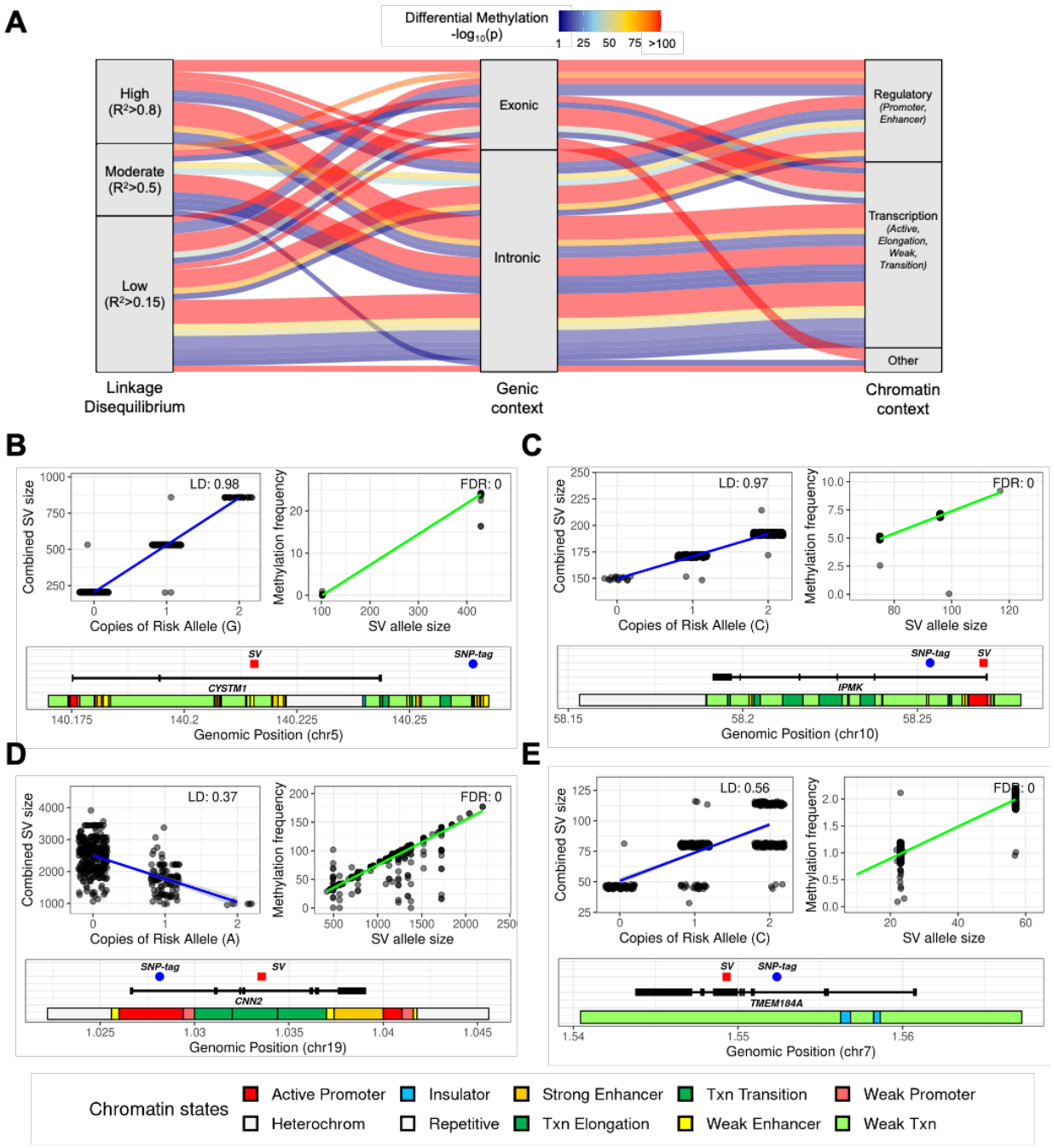


Figure 3: Prioritization of candidate SVs contributing to disease mechanism. **A.** Alluvial plot of the SVs classified as tier 1 across LD level (wrt the AD-haplotype), genic and chromatic contexts. All SVs were significantly associated with allele-specific methylation profiles. **B.** Region surrounding *CYSTM1* gene, in which we identified an intronic AluYb8 (red square, bottom plot) in full LD with the intergenic AD-associated SNP (blue circle, bottom plot). The SNP risk allele (G) co-segregates with the expanded version of the Alu element (top left plot), which is strongly associated with increased local methylation (top right plot). **C.** Region surrounding *IPMK* gene, in which we identify a SV in complete

349 LD with the top AD-associated SNP. The SNP risk allele (G) co-segregates with the expanded version of the Alu
350 element (top left plot), which is strongly associated with increased local methylation (top right plot). **D.** Region in *CNN2*
351 gene, in which an intronic large tandem repeat was found in low LD with the top SNP. The tandem repeat is highly
352 multiallelic, with expanded versions associating with increased AD-risk and bringing increased local methylation. **E.**
353 Region inside *TMEM184A*, in which we identified an exonic SV in moderate LD with the top AD-associated SNP. The
354 SV is biallelic and overlaps a transcriptionally active region based on known chromatin marks. The SNP risk allele (C)
355 is in LD with an expanded version of the SV, which increases local methylation signal.
356
357 Table 1: Structural variants in high LD with AD haplotypes likely contributing to molecular mechanisms.

SNP	**Structural variant**							
Haplotype	*SV*	*LD*	*Size*	*Location*	*Chromatin*	*Diff. Meth*	*Risk*	*Gene*
rs2790214	Unknown	0.97	75-117	Intronic	Promoter	p<5e-150	Larger	*IPMK*
rs7312205	(GCA)n	0.93	51-54	Intronic	Txn	p=5.2e-5	Shorter	*IQCD*
rs442495	(TA)n	0.83	92-157	Intronic	Txn	p=2.5e-81	Larger	*ADAM10*
rs2289702	AluSx1	0.86	52-60	Intronic	Promoter	p=0.001	Larger	*CTSH*
rs12716979	Unknown	0.91	39-46	Partly exonic	Promoter	p<5e-150	Larger	*STX1B*
rs60073596	SVA	0.97	129-167	Intronic	Txn	p<5e-150	Shorter	*TOM1L2*
rs7218117	Unknown	0.93	375-431	Intronic	Txn	p<5e-150	Larger	*MINK1*
rs2154481	AluSx3	0.88	143-156	Intronic	Promoter	p=1.4e-5	Shorter	*APP*
rs148548679	Unknown	0.93	44-51	Partly exonic	Promoter	p<5e-150	Shorter	*ADAM17*
rs79966407	AluSp	0.92	34-57	Intronic	Txn	p=2.4e-6	Larger	*PPP2R3A*
rs6891922	AluYb8	0.98	102-429	Intronic	Txn	p<5e-150	Larger	*CYSTM1*
rs801182	AluSq	0.88	48-52	Intronic	Promoter	p<5e-150	Larger	*VTRNA1*
rs4131290	(AC)n	0.86	61-119	Intronic	Txn	p<5e-150	Shorter	*LMAN2*
rs5011439	AluYb8	0.98	30-363	Intronic	Heterochr	p<5e-150	Larger	*TMEM106B*
rs7072204	(AT)n	0.61	62-80	Intronic	Repressed	p=0.01	Shorter	*PLEKHA1*
rs113187568	AluSx	0.52	45-65	Intronic	Txn	p=0.04	Shorter	*PIK3AP1*
rs383027	(CCCGCG)n	0.52	105-143	Intronic	Txn	p<5e-150	Shorter	*DPEP1*
rs146335962	(CCCG)n	0.65	96-103	Exonic	Promoter	p=7.6e-92	Larger	*DACT3*
rs4379692	(TG)n	0.59	60-78	Partly exonic	Txn	p=0.01	Larger	*NDUFS2*
rs12048591	(AC)n	0.62	58-80	Intronic	Enhancer	p=3.4e-39	Shorter	*EIF4G3*
rs7557129	(TG)n	0.70	51-95	Intronic	Txn	p=0.02	Larger	*MGAT5*
rs111340627	Unknown	0.71	245-374	Intronic	Enhancer	p=3.2e-63	Shorter	*SPON2*
rs335480	Unknown	0.73	0-47	Intronic	Txn	p=7.8e-4	Larger	*DOK3*
rs7384878	(AT)n	0.51	215-310	Intronic	Txn	p<5e-150	Larger	*STAG3LP5*
rs4257937	Unknown	0.55	10-57	Partly exonic	Txn	p<5e-150	Shorter	*TMEM184A*

rs11780835	Unknown	0.53	319-460	Intronic	Txn	p<5e-150	Shorter	*GRINA*

358 *Haplotype*: Most significant AD SNP tagging the haplotype; *SV*: Structural variant type; *LD*: Linkage Disequilibrium
359 between the AD SNP and the SV; Size: minimum and maximum size of the alleles observed in long-read sequencing
360 data; Location: annotation of the SV based on genic context; *Chromatin*: annotation of the SV based on chromatin
361 marks; *Diff. Meth:* p-value of the allele-methylation specific association; *Risk*: whether a larger SV size increases AD
362 (Larger) or not (Shorter); *Gene*: closest gene to the SNP. Note: for visibility purposes, only SV-SNP pairs with LD>0.5
363 and annotated as Tier 1 are shown here. The full table is available in the Supplementary Tables.
364

365 Fine-mapping analysis of a novel haplotype in the *SHARPIN/PLEC* locus

366 We identified a novel AD-associated haplotype at the *SHARPIN/PLEC* locus marked by rs11780835
367 (chr8:143987994:C:A, Figure 4A). Prior GWAS highlighted two missense variants in *SHARPIN*
368 (rs34674752 and rs34173062), where the A alleles (minor allele frequencies 5% and 8%, respectively)
369 increase AD-risk. In contrast, we found that the rs11780835-*G* allele, located ~125 kb upstream of
370 *SHARPIN* and <15 kb upstream of *PARP10* and *PLEC* (sense-strand), is independently associated with
371 AD-risk (GWAS p=2.95e-10; conditional analysis p=2.2e-3, Figure 4A). This allele rs11780835-*G* is both
372 an eQTL and splicing-QTL for *PLEC* across multiple brain regions (cerebellum, cortex, frontal cortex,
373 hippocampus, hypothalamus, and cerebellar hemisphere, based on GTEx), consistently reducing its
374 expression [30]. This aligns with recent reports linking *PLEC* loss to AD, as diminishing its encoded protein,
375 plectin, is associated with AD-reactive glia, promotes tau accumulation, and may contribute to cognitive
376 decline [33, 34].

377 Fine-mapping of the rs11780835 haplotype identified 15 SVs, of which two showed strongest LD with
378 rs11780835: (*i*) a bi-allelic intronic INDEL (LD R^2 = 0.53) involving differential presence of 140 bp, where
379 carriers of rs11780835-*G* frequently carried the deletion (Figure 4D); and (*ii*) a CpG-rich multi-allelic tandem
380 repeat (LD R^2 = 0.30) ranging from 0.18-1.1 kb, in which expanded alleles were associated with increased
381 methylation (Spearman Correlation=0.9, p<2.2e-16), and rs11780835-*G* carriers often harbored the
382 expansions (Figure 4C). Both SVs scored as tier 1 according to our prioritization scheme, and both co-
383 localize with microglia- and oligodendrocyte-specific enhancers, suggesting they confer cell-type-specific

regulation of *PLEC* (Figure 4B, see Methods). While the regulatory impact of the INDEL remains unclear, we hypothesize that expansions in the tandem repeat may reduce *PLEC* expression by methylation-mediated enhancer inactivation, consistent with bulk-tissue eQTL effects of rs11780835-G. Supporting this, single-cell eQTL analysis in the ROS/MAP dataset (see Methods) revealed that rs11780835-*G* carriers exhibit significantly reduced *PLEC* expression across cell types, with the strongest effect in microglia (-0.5 average change in rank-normalised expression; FDR-adjusted q <= 4.67e-6, Figure 4E) [35].

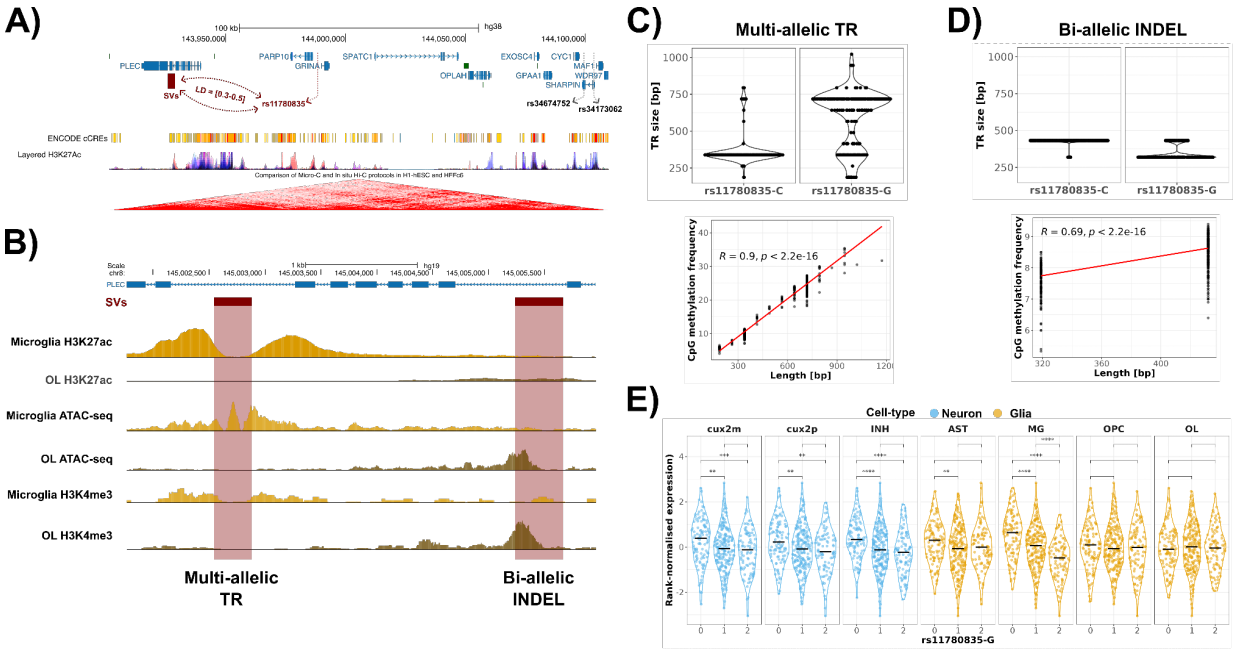


Figure 4E. A novel AD-associated haplotype in the *SHARPIN* locus may be mediated by intronic SVs that confer cell-type-specific regulation of the *PLEC* gene. **A.** UCSC genome browser showing the *SHARPIN* locus with its two well-known AD-associated SNPs, rs34674752 and rs34173062, and the novel independent haplotype marked by rs11780835. The two SVs with the strongest LD (R^2=0.3-0.5) are both intronic in *PLEC*. **B.** Zoomed-in version of *PLEC* (top blue line, exons as blue rectangles), showing the intronic locations of the two SVs (maroon rectangles), along with microglia and oligodendrocyte tracks showing co-localization of short-read-based H3K27ac and ATAC-seq peaks (gold-coloured tracks). **C-D.** Top: length-distributions (Y-axes) in rs11780835-C and rs11780835-G carriers based on long-read phasing. Bottom: average methylation frequency (Y-axes) per allele sequence per genome (black dots) as a function of allele-length (X-axes). **E.** Single-cell eQTL analysis based on ROS/MAP data showing rank-normalised RNA expression (Y-axis), dosage of rs11780835-G allele, across different cell-types (blue are neurons, yellow glia).

402

403 Imputation of SVs from genotyping data

404 We explored the feasibility of imputing SV sizes in individuals with SNP genotyping data. Specifically, SV
405 size was imputed in 2,381 AD cases and 3,555 healthy controls using a random forest model with 5-fold
406 cross-validation (Figure 5A and Methods). Given a haplotype with SNPs and SVs in LD with each other,
407 the model leveraged all SNPs in the haplotype to predict SV size. When an individual SV was found in LD
408 with multiple haplotypes, we used all SNPs from the different haplotypes as predictors. For each SV, the
409 model was trained on 493 individuals with observed SV sizes from long-read sequencing data, and then
410 applied to individuals with genotyping data only. After imputation, 1,651/2,000 (83%) SVs met an imputation
411 quality threshold of R^2>0.5. Imputation quality was generally high (median=0.77, IQR=0.3, Figure 5B, Table
412 S7) and was positively associated with the LD between the SV and the most significant SNP in the
413 haplotype (Spearman's correlation=0.28, p=2.5e-30), indicating better performance for SVs more tightly
414 linked to nearby SNPs.
415
416 With this larger sample size (2,381 AD cases and 3,555 healthy controls), we performed association
417 analysis between AD status and AD-haplotypes (*i.e* SNPs, Figure 5C). As expected, we detected multiple
418 known AD risk haplotypes, with the strongest signals being those near *BIN1* (rs6733839, p=1.9e-6,
419 OR_C=0.83, 95% CI=[0.77-0.90]), *PLCG2* (rs12444183, p=1.7e-5, OR_A=0.85 [0.79-0.91]), and *PICALM*
420 (rs10792832, p=2.0e-5, OR_A=0.85 [0.79-0.91]) (Table S8). Using the same multiple-testing correction
421 factor, we also tested imputed SV sizes for association with AD status. In total, 112 SVs significantly
422 associated with AD risk (Table S9), with several examples including: 4 correlated SVs near *PICALM*, the
423 most significant of which was the intronic SINE element of the Alu family chr11:86,036,934-86,036,969
424 ($size_{REF}$=35bp, OR=1.02 [1.01-1.03], p=1.17e-5, imputation R^2=0.93, AD-haplotype: rs10792832 with LD
425 R^2=0.60, lying in a transcription-related chromatin region, tier 3); an intronic tier 1 Alu element
426 chr16:23,606,810-23,606,851 in/near *NDUFAB1* ($size_{REF}$=41, OR=0.97 [0.95-0.98], p=4.7e-5, imputation
427 R^2=0.84, AD-haplotype: rs35228710 with LD R^2=0.20, located in a transcription-related chromatin region);
428 another intergenic Alu element chr17:63,472,601-63,472,648 in *ACE* ($size_{REF}$=47, OR=0.97 [0.95-0.98],

429 p=5.9e-5, imputation R^2=0.81, AD-haplotype: rs4277405 with R^2 LD=0.8, located in a heterochromatin
430 region, tier 3) (Table S1).
431 We finally examined whether modeling multiple AD-associated haplotypes and/or SVs provided a better fit
432 to AD status than testing individual AD-associated SNPs alone. To this end, we compared a series of
433 logistic regression models with elastic net regularization in the same 2,381 AD cases and 3,555 healthy
434 controls (see Methods). Model fit was assessed using likelihood ratio tests relative to a baseline model
435 including only population structure covariates. Across loci, we observed heterogeneity in the model
436 configurations that provided the best fit. In a minority of regions, associations were adequately captured by
437 individual SNPs (*PICALM, IGFL1, BLNK*). In contrast, for a larger fraction of loci, jointly modeling multiple
438 SNPs tagging independent haplotypes resulted in improved model fit, including well-established regions
439 such as *BIN1, PLCG2, HLA,* and *MS4A6A* (Figure 5D, Table S10). At additional loci, models incorporating
440 multiple SVs in LD with haplotypes provided comparable or better fit than SNP-based models alone, e.g
441 near *NDUFAB1* (which comprises 33 SVs including 8 tier 1 in moderate-to-low LD), *NOVA2* (including tier
442 1 SVs in moderate LD), and *CD33* (including tier 2 SVs in low LD). Finally, in a subset of loci including *ACE,*
443 *RHOH, APH1B*, joint models including both SNPs and SVs yielded the best fit, indicating that both types of
444 variation may contribute to the observed association signal.
445
446 Overall, these analyses primarily illustrate the feasibility of imputing SV sizes from SNP-genotype data, and
447 that testing SVs as well as the joint-effect of multiple haplotypes in a region may contribute to refining
448 association signals.

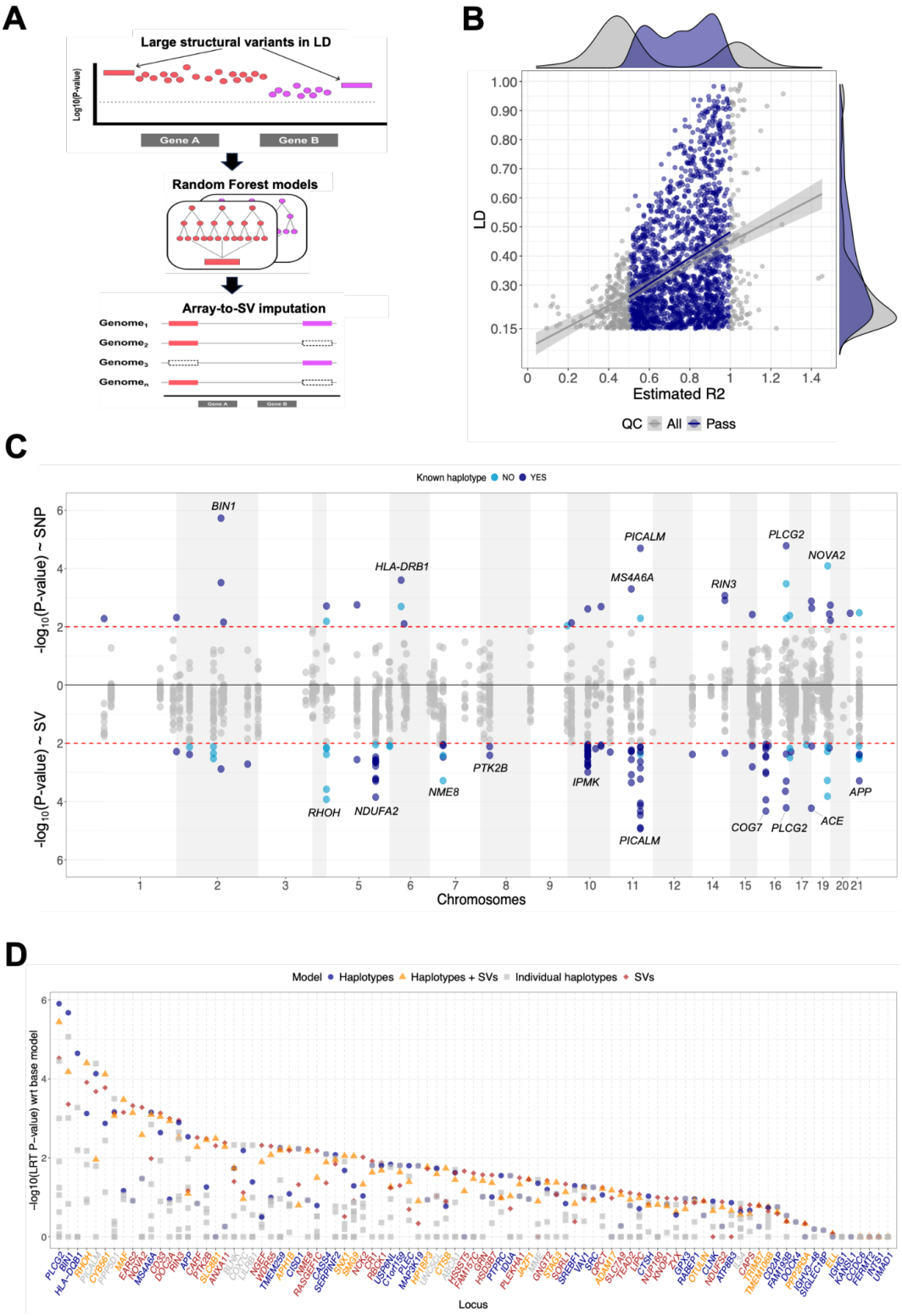


Figure 5: Imputation of SV sizes using SNP genotypes and association analyses. **A.** Schematic representation of the imputation workflow. In short, we used all SNPs in the same haplotype as the SV to predict SV size. As a model, we

used Random Forest models trained and evaluated using 5-fold cross-validation (cv=5), training in the individuals with measured SV size based on long-read sequencing data, and testing on the individuals with only SNP array data available. **B.** Correlation and marginal densities between imputation quality and LD between the SV and the AD-haplotypes, in *all* SVs as well as the 1,381 SVs after QC (R^2>0.6). We found a significant, positive correlation between LD and R^2 (Spearman's correlation=0.27, p=2.5e-30) **C.** Miami plot of the association with AD of imputed SNPs and SVs. The associations are based on the comparison of 2,381 AD cases as opposed to 3,555 healthy controls. Highlighted the significant associations after correction for multiple testing (FDR>10%). As correlation exists between SVs, we used the same correction factor as the SNP comparison. Dark dots refer to SNPs and SVs in LD with known haplotypes, while light dots refer to significant association in novel haplotypes. The top genes are annotated, for both SNPs and SVs. **D.** Comparison of association statistics between a base model including only PC1-5, and (i) individual haplotypes (yellow squares), (ii) combined haplotypes (blue circle), and (iii) combined haplotypes plus SVs (red triangles). For the models, we used elastic net regressions with Lasso regularization. To compare models with the base model, we used likelihood ratio tests.

Discussion

In this study, we leveraged public GWAS summary statistics, long-read sequencing and epigenomic data to systematically characterize structural variants (SVs) within Alzheimer's disease (AD) risk haplotypes. We show that 74% of AD-haplotypes are in linkage disequilibrium (LD) with one or more SVs. This adds to growing evidence that SVs co-occur within AD risk haplotypes and could contribute to disease mechanisms. By integrating allele-specific methylation, genomic localization, and chromatin features, we further prioritized a subset of SVs with convergent regulatory evidence. Finally, we show that these SVs can be imputed from SNP array data, enabling their evaluation in GWAS-scale cohorts. Consistent with this, at several loci we showed that jointly modelling multiple haplotypes and/or SVs provided a better fit to AD risk than single SNP associations alone, in our cohort.

Among approximately one million SVs detected genome-wide, we identified 2,000 SVs co-occurring with AD risk haplotypes. These SVs are predominantly multiallelic, include mostly transposable elements (TE)

479 insertions and tandem repeats (TRs), and are enriched in enhancer- and transcription-related chromatin
480 marks. Importantly, only a subset of these SVs showed allele-specific differential methylation and
481 localization within genes and regulatory or transcriptionally active regions. We stratified SVs into three tiers
482 of regulatory relevance, revealing that approximately 25% of AD-haplotypes harbor at least one tier 1 SV.
483 These findings suggest that structural variation may represent a recurrent regulatory mechanism across
484 AD loci, rather than isolated examples at individual genes. To further support this, we observed that AD-
485 haplotypes harboring SVs were twofold more likely to be associated with changes in gene expression levels
486 (*i.e.* eQTLs) than haplotypes without SVs.
487
488 We identified significant allele-specific methylation differences for 18% of SVs, and SV-based methylation-
489 QTL signals were consistently stronger than those of nearby SNPs. This supports a model in which
490 multiallelic structural variants modulate local chromatin configuration and DNA methylation, potentially
491 influencing gene expression. For example, in several loci including *SHARPIN/PLEC*, *TMEM106B*,
492 *TMEM184A, IPMK*, and others, expanded SV alleles were associated with marked increased in methylation
493 within enhancers or transcriptionally active regions, which could explain mechanistically the underlying
494 GWAS signal. Interestingly, as exemplified by the fine-mapping of *PLEC* locus, such regulatory
495 mechanisms could be cell- and tissue-specific.
496
497 Our findings align with and extend previous reports of SV involvement at AD loci. For example, we detected
498 two tier 1 SVs at *TPCN1* and *TMEM106B* loci (in LD with AD-haplotypes rs7312205 and rs5011439,
499 respectively), that involve complete deletions of Alu-elements (~320 bp, Alu-Y subfamily), consistent with
500 recent reports [23, 36, 37]. We also observed a known Alu-based 289 bp insertion (scored as tier 2) in an
501 intron of the *ACE* gene, located ~17 Kbp downstream of rs4277405 (moderate LD, R^2=0.33), which was
502 previously documented in AD [38]. Similarly, our data confirmed moderate-to-strong LD between SVs and
503 AD-haplotypes at the *NBEAL1*, *JAZF1*, and *APP* loci. These include tier 1 and tier 2 TE insertions and
504 tandem repeats associated with AD-haplotypes tagged by rs115600411, rs67250450, and rs2154481,
505 respectively [13]. In addition to validating these known SNP-SV relationships, we identified several novel
506 SV events at AD risk loci that have not been previously described, such as the tier 1 TE and TR in/near

507 *PLEC*, *IPMK, LMAN2, CNN2*, and others. Importantly, our work moves beyond cataloging LD relationships
508 by systematically integrating epigenetic context to distinguish likely functional variants. These new SVs
509 provide fresh candidates for functional studies and further emphasize that the genetic architecture of AD
510 loci is more complex than captured by SNPs alone.
511
512 Beyond SV prioritization, our analyses also revealed substantial allelic heterogeneity at known AD loci in
513 terms of the number of haplotypes and the number of SVs. Through conditional analysis (COJO), we
514 identified 280 independent haplotypes across 98 loci, with most loci harboring multiple independently
515 associated haplotypes. By design, we included sub-significant haplotypes that did not reach genome-wide
516 significance in the latest meta-analyses, yet they might still carry meaningful disease risk.
517
518 While the prioritization framework identifies SVs with strong regulatory support, evaluating their association
519 with AD risk requires larger cohorts. Because SVs are typically absent from standard GWAS datasets and
520 imputation panels, we explored the feasibility of SV imputation from SNP-array data using a random forest
521 model. This approach enabled the imputation of 83% of SVs with good accuracy (median R^2=0.77) and
522 allowed association testing in ~6,000 individuals. In this larger cohort, we identified 112 imputed SVs
523 significantly associated with AD risk, including those in *BIN1, PICALM, SORL1*, and *ACE.* Moreover, model
524 comparison analyses demonstrated that in the majority of loci, multi-variant models incorporating multiple
525 haplotypes and/or SVs provided a better fit to AD status than single-SNP models alone. In several regions,
526 models including SVs alone or combined SNP-SV models outperformed SNP-only models, highlighting that
527 structural variation can contribute independently or jointly to the observed association signals. Although
528 imputation accuracy correlated with LD, we observed that multi-SNP signatures could tag certain multiallelic
529 SVs even at moderate or low LD levels. This demonstrates the possibility of incorporating structurally
530 complex variation into large-scale association studies. Further methodological advances, particularly
531 incorporating phasing (similar to traditional SNP imputation methods such as BEAGLE and Minimac) [39,
532 40] and larger long-read reference panels, will be required to improve the recall of SVs in cohort-wide
533 studies.
534

535 Several limitations warrant consideration. The long-read discovery cohort remains modest relative to
536 GWAS-scale studies, limiting the detection of low frequency and rare SV alleles, and potentially affecting
537 imputation performance. Our analysis was restricted to individuals of European (specifically Dutch)
538 ancestry, and structural variant architecture differs significantly across populations [41]. Additionally, while
539 allele-specific methylation and chromatin annotation provide evidence for regulatory involvement, functional
540 validation will be required to establish direct causal mechanisms.
541
542 Despite these caveats, our findings highlight the importance of integrating SVs into genetic models of
543 complex disease. By combining haplotype structure, long-read SV discovery, methylation profiling, and
544 chromatin annotation, we demonstrate that many AD-associated loci harbor regulatory SVs that may
545 underlie GWAS signals. Rather than viewing GWAS loci as single-SNP signals, our results reinforce a
546 model in which multiallelic SVs, embedded within complex haplotypes, perturb local epigenetic landscapes,
547 and contribute to gene dysregulation in AD.
548
549 In summary, structural variants are not peripheral to AD genetics but represent an integral component of
550 its architecture. Systematic prioritization using epigenomic context enables distinction between merely
551 linked variants and those with plausible regulatory impact. Future work integrating long-read sequencing,
552 improved SV imputation, and cell-type-specific functional assays will be critical to refine causal inference
553 and translate these findings into actionable mechanistic insight.
554

555 Methods

556 Genotyping data

557 We included 5,936 individuals with available genotyping data in this study: (i) 346 cognitively healthy
558 centenarians from the 100-plus Study [28]; (ii) 83 older adults from the 100-plus Study, corresponding to
559 the partners of the centenarians' offspring [28]; (iii) 1859 patients clinically diagnosed with probable AD

560 from the Amsterdam Dementia Cohort (ADC) [29]; (iv) 422 pathologically confirmed AD patients from the
561 Netherlands Brain Bank (NBB) [42]; (v) 1651 Dutch older adults from the Longitudinal Aging Study of
562 Amsterdam (LASA) [43, 44]; (vi) 1276 healthy older from the ADC [29]; (vii) 57 controls from the NBB[42];
563 (viii) 98 individuals from the Dutch twin registry [45]. This dataset of individuals is genetically unrelated and
564 of European ancestry. A detailed description of the individuals included, their respective cohorts, and the
565 quality control of the genetics data is available elsewhere [46].
566

567 Selection and association of AD variants from GWAS

568 We focus on N=118 SNPs identified to have a genome-wide significant association with AD from previous
569 AD-GWAS. From the GWAS consortium, we obtained the summary statistics of these SNPs as well as all
570 SNPs lying within 250 kbp (kilo base pairs) from them. Pairwise linkage disequilibrium (LD) values between
571 SNPs were calculated using genotyping data from the 5,936 individuals with available genotyping data. We
572 performed a clumping analysis with plink2 (v2.00a6LM) [47] using a R^2 threshold of 0.15, a default window
573 of 250 kbp, and keeping SNPs with a p-value of association <5e-5. This allowed us to identify the major
574 common haplotypes (each represented by the most significant SNP) with suggestive evidence of
575 association with AD, based on LD structure from our genotyping data.
576 To further refine the list of haplotypes, we performed a joint effects analysis with GCTA-COJO (v1.94.1)
577 [48], specifying a stepwise model (--cojo-slct), a minor allele frequency threshold of >2.5%, and a joint-
578 effect p-value of association <0.05 (default option). This procedure jointly re-estimates SNP effects by
579 conditioning on other variants in LD, effectively yielding to a set of (semi-)independent haplotypes (each
580 represented by a single SNP) with a significant association on AD.
581

582 Long-read sequencing data and processing

583 We included N=493 individuals with long-read sequencing data produced with the PacBio Sequel2
584 instrument. Of these, 245 are patients diagnosed with Alzheimer's disease from the Amsterdam University

585 Medical Center, while the remaining N=248 are cognitively healthy centenarians from the 100-plus Study
586 cohort. Additional information regarding the samples included, the respective cohorts, and the long-read
587 sequencing data generation and quality assessment is available elsewhere [21, 41, 46]. Briefly, genomic
588 DNA from the blood of all individuals underwent strict quality assessment to ensure its integrity before
589 sequencing [21, 41]. After sequencing, raw reads were collected and analyzed through an in-house pipeline
590 (https://github.com/holstegelab/snakemake_pipeline). This pipeline generates both high quality HiFi data
591 as well as lower quality non-HiFi data for each sample, and aligns data to the reference genome (GRCh38,
592 patch release 14). When available, individual BAM-files from the same sample were merged into a single
593 BAM-file.
594 The Medical Ethics Committee (METC) of the Amsterdam UMC approved all studies for which genetics
595 data was included in this manuscript, both for the genotyping data and the long-read sequencing data. All
596 individuals and/or their legal representatives provided written informed consent for participating in clinical
597 and genetic studies.
598

599 Structural variant calling and repeat annotation

600 We identified structural variants in the long-read sequencing data using the same procedure detailed in
601 Salazar et al. 2025 [23]. Briefly, we first identified candidate SVs using sniffles2 (v2.0.7) [49] with default
602 parameters. Candidate SVs coordinates were then converted to a BED-file format. SVs annotations
603 including Simple_repeat, Low_complexity, and Satellite were downloaded from the UCSC genome browser
604 [24], and their coordinates were merged with the location of the above candidate SVs. We used otter (v1.0)
605 [41] to perform local assembly of each region in each sample. As HiFi data alone was shown to be sensitive
606 to allele dropout events especially in large SVs, we performed local assembly using both HiFi and non-HiFi
607 data (read quality >0.85) [41]. We joint-genotyped all allele-sequences per SV with otter, generating a SV-
608 based gVCF across all individuals. We removed SVs overlapping centromere regions and known segmental
609 duplications based on annotations reported in the UCSC Genome Browser. Finally, we restricted to SVs
610 lying within 250kbp of the 84 AD-associated SNPs from GWAS. In addition, we excluded SVs (i) that were

611 not called in >10% of individuals, (ii) with a local-to-global coverage ratio <0.3 or >3.3 [41], and (iii) with an
612 allelic coverage <3 (i.e, less than 3 sequencing reads supporting each allele sequence).
613

614 Linkage analysis of SNPs and SVs

615 To estimate LD between the joint-genotypes of all SVs and SNPs, we used both quantitative-trait-locus
616 (QTL) approach as well as a classic LD approach. The QTL analysis is typically used to model the
617 relationship between SNPs and quantitative traits, e.g., gene expression levels, blood pressure, or height.
618 Given that SVs are often multiallelic, we considered the size of SVs as the quantitative trait. Thus, we tested
619 the association between SV sizes and SNP dosages of all SNPs lying within 250kb up/downstream of the
620 SV location. For the association, we used linear regression models while adjusting for population
621 stratification (PC 1-5). QTL analysis was performed with plink2 (v2.00a6LM) [47].
622 In addition to the QTL approach, we calculated classic LD between SNPs and SVs by scaling the size of
623 each SV within the 0-2 range. This allowed us to treat SVs as SNP dosages, as they also range 0-2, and
624 to calculate LD accordingly with plink2. We assessed the robustness and concordance of the two
625 approaches by correlating the QTL-p-value (in -log10 scale) with the LD values. Then we kept SNP-SV
626 relationships with QTL-p-value <5e-8 and LD>0.15.
627 We finally combined this dataset with the independent haplotypes significantly associated with AD after the
628 GCTA-COJO analysis [48], thus linking AD-associated haplotypes with SVs in LD with them.
629

630 Methylation analysis

631 Allele-specific methylation profiles were derived using haplr (https://github.com/holstegelab/haplr), a
632 probabilistic framework that assigns long-reads from BAM files to predefined haplotypes using a Bayesian
633 model based on haplotype-informative variants. Briefly, for each read, haplr computes posterior
634 probabilities across competing haplotypes by integrating allele matches at informative sites while
635 accounting for sequencing error probabilities. Reads are assigned to the most likely haplotype when

636 predefined confidence criteria are met. Methylation calls embedded in the BAM files are aggregated per
637 haplotype to obtain haplotype specific CpG coverage, methylation counts, and methylation fractions. To
638 identify SV alleles differentially associated with methylation, we then modeled methylation levels as a
639 function of SV allele size using linear regression (methylation ~ allele size). Resulting p-values were
640 corrected for multiple testing using False Discovery Rate (FDR) and considered significant at FDR<0.05.
641
642 In parallel, we performed cis-methylation quantitative trait locus (mQTL) analyses to compare the regulatory
643 impact of SVs and SNPs within the same loci. For each SV, CpG sites located within ±250 kb were
644 considered, requiring methylation measurements in at least 25% of individuals. For each CpG, methylation
645 levels were then tested for association with SV length or genotype dosage (for SNPs) using robust linear
646 regression models while adjusting for population stratification (PC1-5). Associations were considered
647 significant at FDR<0.05. For loci where SNPs and SVs were in LD, we directly compared the strength of
648 association signals between variant classes.
649

650 Prioritization of SVs

651 SVs were prioritized using a composite functional annotation score integrating epigenetic and genomic
652 context. Four components were considered: (i) differential methylation status (FDR ≤ 0.05 vs. non-
653 significant vs. missing), (ii) genic context (genic vs. intergenic), (iii) genic annotation (exonic, partly exonic,
654 or intronic), and (iv) chromatin state (promoter/enhancer vs. transcription/insulator/repressed vs.
655 heterochromatin/repeat vs. unannotated). Each component was assigned an ordinal priority score (lower
656 values reflecting higher functional relevance). Differential methylation was weighted two-fold to emphasize
657 functional epigenetic evidence. A final priority score was calculated as the weighted sum of all components,
658 and SVs were ranked accordingly. Tier boundaries were defined based on the empirical distribution of the
659 composite score (range 5-15, median=12, mean=11.7), with tier 1 defined as scores ≤ 8, tier 2 as scores
660 9-10, and tier 3 as scores > 10.

661 # Fine-mapping analysis of *SHARPIN/PLEC* locus

662 We first evaluated all regulatory-candidate SVs in the SHARPIN locus, by filtering SVs that were ≥ 25 bp
663 and were at least in low LD ($R^2 \geq 0.2$) with rs11780835. Two candidate SVs remained, both of which were
664 intronic in PLEC. We performed targeted long-read phasing to evaluate how carriership of rs11780835-G
665 allele associated with the *haplr* (https://github.com/holstegelab/haplr): for each genome, we extracted
666 GWAS imputed haplotypes within the *PLEC* locus, determined rs11780835-allele status, and assigned
667 read-membership to each haplotype based on spanning read-alignments of nearby heterozygous common
668 SNPs; enabling us to determine haplotype-level SV-length per rs11780835-allele. Similarly, for each read-
669 set per haplotype per genome, we extracted all methylated CpG sites that corresponded to the allele-
670 sequence of each SV and calculated the average number of methylated CpGs. To minimise noise, we
671 considered a CpG site as 'methylated' if it had a based modification probability (e.g., 'ML' tag in BAM/SAM
672 file) of at least 50%. To co-localize the SVs with cell-type-specific enhancers, we performed a 'lift-over'
673 procedure of the SV-coordinates from HG38 to HG19 via the UCSC genome browser and overlapped
674 chromatin profile peaks from Nott et al. 2019 [50]. Finally, we performed an cell-type-specific eQTL analysis
675 between dosage of rs11780835-G and *PLEC* gene expression using single-nucleus RNA-sequencing data
676 from ROS/MAP using the same procedure outlined in Salazar et al. 2025 which is based on the 'd-sum'
677 normalisation procedure from Cuomo et al. 2021 [23, 51].
678

679 # Imputation of SVs based on genotyping data

680 We imputed SV sizes using a random forest model with 5-fold cross validation. For each SV, we used
681 nearby SNPs in linkage disequilibrium (LD) as predictors. Specifically, let SV_i be an SV in LD with an AD-
682 associated haplotype (H_j), represented by the top SNP, SNP_j. We selected as features all $SNP_1, SNP_2, \ldots,$
683 SNP_n in LD ($R^2 > 0.15$) with SNP_j to predict SV_i size. The random forest model was trained on 493
684 individuals with long-read sequencing data and applied to 5,936 individuals with genotyping data. To assess
685 imputation quality, we computed an imputation quality score (R^2) analogous to SNP imputation quality

686 scores (e.g., BEAGLE) [39], defined as the ratio of the expected variance in the training dataset to the
687 observed variance in the test dataset. For downstream analyses, we kept SVs with imputation $R^2>0.5$.
688 After imputation, we performed a SV-GWAS analysis by comparing the size of each SV between 2,381 AD
689 cases and 3,555 healthy controls. In the same sample set, we also performed a SNP-GWAS analysis by
690 comparing the dosages of AD-haplotypes. In both cases, we used logistic regression models, correcting for
691 population stratification (PC1-5).
692

693 Comparison of SNP-based and haplotype-based associations

694 We evaluated whether the inclusion of multiple haplotypes (each represented by a lead SNP) and structural
695 variants (SVs) enhances the association with AD status compared to models considering individual SNPs
696 alone. Let $locus_i$ denote an AD-associated locus identified in the original GWAS, for which our COJO
697 analysis identified n_{haplo} independent and significant haplotypes ($hap_1, hap_2, \ldots, hap_{n\text{-}haplo}$). Additionally, for
698 each $locus_i$, we identified n_{SV} SVs ($sv_1, sv_2, \ldots, sv_{n\text{-}sv}$) in LD with these haplotypes. Using 2,281 AD cases
699 and 3,511 healthy controls, we first established a baseline model including only the top five principal
700 components (PC1-5). We then extended this model in three steps: (1) *single* haplotype model, incorporating
701 each haplotype ($hap_1, hap_2, \ldots, hap_n$) separately; (2) *combined* multiple-haplotype model, including all
702 haplotypes simultaneously to capture their combined additive effects; and (3) a *complete* model, further
703 extending the multiple-haplotype model by incorporating the SVs ($sv_1, sv_2, \ldots, sv_{n\text{-}sv}$). Each model was
704 trained using a logistic elastic net regression framework, implemented via the H2O package in R [52], to
705 enforce sparsity and reduce collinearity between predictors, particularly in the *complete* model. Each model
706 was then compared to the baseline model using likelihood ratio tests (LRTs), where the LRT p-value
707 evaluated whether the inclusion of additional genetic components significantly improved the model's
708 predictive power upon the baseline model.
709

Competing interest statement

The authors have declared no competing interest.

Data availability

All data produced in the present study are available upon reasonable request to the authors.

Funding

This work was supported by Scientific Excellence program 2014 (IPB), Stichting Alzheimer Nederland (WE09.2014-03), Stichting Dioraphte (VSM 14 04 14 02) and Stichting VUmc Fonds, Hans und Ilse Breuer Stiftung Alzheimer Research Prize 2020. H. Holstege and M.J.T. Reinders are recipients of ABOARD, a public-private partnership receiving funding from ZonMW Nationaal Dementiaprogramma (#73305095007) and Health~Holland, Topsector Life Sciences & Health (PPP-allowance; #LSHM20106). More than 30 partners, including de Hersenstichting (Dutch Brain Foundation) participate in ABOARD (https://www.alzheimer-nederland.nl/onderzoek/projecten/aboard).

References

1. 2012 Alzheimer's disease facts and figures. Alzheimer's & Dementia. 2012;8:131–168.
2. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, et al. Role of genes and environments for explaining Alzheimer disease. Archives of General Psychiatry. 2006;63:168–174.
3. Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet. 2022;54:412–436.
4. ADGC, Bonn, CHARGE, EADB, Bellenguez C, EADI, et al. Consensus meta-analysis of genome-wide association studies for Alzheimer's disease and related dementia. 2025.
5. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–1162.
6. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022;19:823–826.

7. Klein SJ, O'Neill RJ. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. Chromosome Res. 2018;26:5–23.
8. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018;19:286–298.
9. McMurray CT. Mechanisms of trinucleotide repeat instability during human development. Nat Rev Genet. 2010;11:786–799.
10. Khristich AN, Mirkin SM. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. Journal of Biological Chemistry. 2020;295:4134–4170.
11. Guo MH, Lee W-P, Vardarajan B, Schellenberg GD, Phillips-Cremins JE. Polygenic burden of short tandem repeat expansions promotes risk for Alzheimer's disease. Nat Commun. 2025;16:1126.
12. On Behalf of the BELNEU Consortium, De Roeck A, Duchateau L, Van Dongen J, Cacace R, Bjerke M, et al. An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. Acta Neuropathologica. 2018;135:827–837.
13. Wang H, Dombroski BA, Cheng P-L, Tucci A, Si Y-Q, Farrell JJ, et al. Structural Variation Detection and Association Analysis of Whole-Genome-Sequence Data from 16,905 Alzheimer's Diseases Sequencing Project Subjects. 2023.
14. Lunnon K, Smith R, Hannon E, De Jager PL, Srivastava G, Volta M, et al. Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. Nat Neurosci. 2014;17:1164–1170.
15. De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci. 2014;17:1156–1163.
16. Smith RG, Pishva E, Shireby G, Smith AR, Roubroeks JAY, Hannon E, et al. A meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex. Nat Commun. 2021;12:3517.
17. Zhang Q, Marioni RE, Robinson MR, Higham J, Sproul D, Wray NR, et al. Genotype effects contribute to variation in longitudinal methylome patterns in older people. Genome Med. 2018;10:75.
18. Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, Miller GE, et al. Factors underlying variable DNA methylation in a human community cohort. Proc Natl Acad Sci USA. 2012;109:17253–17260.
19. Sakharkar KR, Sakharkar MK, Culiat CT, Chow VTK, Pervaiz S. Functional and evolutionary analyses on expressed intronless genes in the mouse genome. FEBS Letters. 2006;580:1472–1478.
20. Zhang F, Gu W, Hurles ME, Lupski JR. Copy Number Variation in Human Health, Disease, and Evolution. Annu Rev Genom Hum Genet. 2009;10:451–481.
21. Salazar A, Tesi N, Knoop L, Pijnenburg Y, Van Der Lee S, Wijesekera S, et al. An AluYb8 retrotransposon characterises a risk haplotype of TMEM106B associated in neurodegeneration. 2023.
22. Cabron A-S, Borgmeyer U, Richter J, Peisker H, Gutbrod K, Dörmann P, et al. Lack of a protective effect of the Tmem106b "protective SNP" in the Grn knockout mouse model for frontotemporal lobar degeneration. Acta Neuropathol Commun. 2023;11:21.
23. Salazar A, Tesi N, Knoop L, Pijnenburg Y, Van Der Lee S, Wijesekera S, et al. Fine-Mapping of the TMEM106B Locus Reveals Four Haplotypes That are Differentially Associated With Risk for Neurodegeneration. 2025.
24. Hinrichs AS. The UCSC Genome Browser Database: update 2006. Nucleic Acids Research. 2006;34:D590–D598.
25. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2016;44:D733-745.
26. Tarailo-Graovac M, Chen N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. CP in Bioinformatics. 2009;25.
27. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–216.
28. Holstege H, Beker N, Dijkstra T, Pieterse K, Wemmenhove E, Schouten K, et al. The 100-plus Study of cognitively healthy centenarians: rationale, design and cohort description. European Journal of Epidemiology. 2018. October 2018. https://doi.org/10.1007/s10654-018-0451-3.
29. van der Flier WM, Scheltens P. Amsterdam Dementia Cohort: Performing Research to Optimize Care. Journal of Alzheimer's Disease. 2018;62:1091–1111.

30. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013;45:580–585.
31. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2019;47:D886–D894.
32. Tesi N, van der Lee S, Hulsman M, Holstege H, Reinders MJT. snpXplorer: a web application to explore human SNP-associations and annotate SNP-sets. Nucleic Acids Research. 2021;49:W603–W612.
33. Koskeridis F, Fancy N, Tan PF, Meena D, Evangelou E, Elliott P, et al. Multi-trait association analysis reveals shared genetic loci between Alzheimer's disease and cardiovascular traits. Nat Commun. 2024;15:9827.
34. Valencia RG, Mihailovska E, Winter L, Bauer K, Fischer I, Walko G, et al. Plectin dysfunction in neurons leads to tau accumulation on microtubules affecting neuritogenesis, organelle trafficking, pain sensitivity and memory. Neuropathology Appl Neurobio. 2021;47:73–95.
35. DE JAGER P, BENNETT D. The Religious Orders Study and Memory and Aging Project. 2015.
36. Kaivola K, Chia R, Ding J, Rasheed M, Fujita M, Menon V, et al. Genome-wide structural variant analysis identifies risk loci for non-Alzheimer's dementias. Cell Genomics. 2023;3:100316.
37. Vialle RA, De Paiva Lopes K, Li Y, Ng B, Schneider JA, Buchman AS, et al. Structural variants linked to Alzheimer's disease and other common age-related clinical and neuropathologic traits. Genome Med. 2025;17:20.
38. Wu S-J, Hsieh T-J, Kuo M-C, Tsai M-L, Tsai K-L, Chen C-H, et al. Functional regulation of Alu element of human angiotensin-converting enzyme gene in neuron cells. Neurobiology of Aging. 2013;34:1921.e1-1921.e7.
39. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. The American Journal of Human Genetics. 2021;108:1880–1890.
40. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;44:955–959.
41. Tesi N, Salazar A, Zhang Y, Van Der Lee S, Hulsman M, Knoop L, et al. Characterizing tandem repeat complexities across long-read sequencing platforms with TREAT and *otter*. Genome Res. 2024;34:1942–1953.
42. Rademaker MC, de Lange GM, Palmen SJMC. The Netherlands Brain Bank for Psychiatry. Handbook of Clinical Neurology, vol. 150, Elsevier; 2018. p. 3–16.
43. Huisman M, Poppelaars J, van der Horst M, Beekman AT, Brug J, van Tilburg TG, et al. Cohort Profile: The Longitudinal Aging Study Amsterdam. International Journal of Epidemiology. 2011;40:868–876.
44. Hoogendijk EO, Deeg DJH, Poppelaars J, van der Horst M, Broese van Groenou MI, Comijs HC, et al. The Longitudinal Aging Study Amsterdam: cohort update 2016 and major findings. European Journal of Epidemiology. 2016;31:927–945.
45. Willemsen G, de Geus EJC, Bartels M, van Beijsterveldt CEMT, Brooks AI, Estourgie-van Burk GF, et al. The Netherlands Twin Register Biobank: A Resource for Genetic Epidemiological Studies. Twin Research and Human Genetics. 2010;13:231–245.
46. Tesi N, Van Der Lee S, Hulsman M, Van Schoor NM, Huisman M, Pijnenburg Y, et al. Cognitively healthy centenarians are genetically protected against Alzheimer's disease. Alzheimer's & Dementia. 2024;20:3864–3875.
47. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics. 2007;81:559–575.
48. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat Commun. 2018;9:224.
49. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024;42:1571–1580.
50. Nott A, Holtman IR, Coufal NG, Schlachetzki JCM, Yu M, Hu R, et al. Brain cell type–specific enhancer–promoter interactome maps and disease - risk association. Science. 2019;366:1134–1139.
51. Cuomo ASE, Alvari G, Azodi CB, single-cell eQTLGen consortium, McCarthy DJ, Bonder MJ. Optimizing expression quantitative trait locus mapping workflows for single-cell studies. Genome Biol. 2021;22:188.

52. Chen Y, Liu X, Gao L, Zhu M, Shia B-C, Chen M, et al. Using the H2O Automatic Machine Learning Algorithms to Identify Predictors of Web-Based Medical Record Nonuse Among Patients in a Data-Rich Environment: Mixed Methods Study. JMIR Med Inform. 2023;11:e41576.