Fine-mapping of the *TMEM106B* locus identifies four haplotypes with differential associations to neurodegeneration

AUTHORS:

Alex Salazar[1], Niccolò Tesi[1,2], Lydian Knoop[1], Martijn Huisman,[3,4] Natasja M. van Schoor,[3,4] Betty Tijms[5,6], Jort Vijverberg[5,6], Sven van der Lee[1,6], Sanduni Wijesekera[1], Jana Krizova[1], Mikko Hiltunen[7], Markus Damme[8], Leonard Petrucelli[9,10], Marcel Reinders[2], August B. Smit[11], Anke A. Dijkstra[12,13], Marc Hulsman[1], ADGC, Bonn, CHARGE, EADB, EADI, FinnGen, GERAD, GR@ACE/DEGESCO, PGC-ALZ, and Henne Holstege[1,3,4,14,15]

AFFILIATIONS

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
7. Institute of Biomedicine, University of Eastern Finland. Kuopio, Finland
8. Institute of Biochemistry, Christian-Albrechts-University Kiel, Otto-Hahn-Platz 9, Kiel, Germany
9. Department of Neuroscience, Mayo Clinic, Jacksonville, FL, USA
10. Neuroscience Graduate Program, Mayo Clinic Graduate School of Biomedical Sciences, Jacksonville, FL, USA
11. Department of Molecular and Cellular Neurobiology, Center for Neurogenomics and Cognitive

Research, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

12. Department of Pathology, Amsterdam UMC, Amsterdam Neuroscience, Amsterdam, Netherlands

13. Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands

14. VIB Center for Brain and Disease Research, Leuven, Belgium

15. Department of Neurosciences, Leuven Brain Institute, KU Leuven, Leuven, Belgium

CORRESPONDING AUTHOR:

Henne Holstege: h.holstege@amsterdamumc.nl

KEYWORDS

TMEM106B, Alzheimer's disease, Fine-mapping, Haplotypes, Long-read sequencing, Structural variation, DNA methylation, Centenarians, Cognitive resilience, Extreme longevity

ABSTRACT

Genetic variation at the *TMEM106B* locus has been associated with Alzheimer's disease (AD) and related neurodegenerative phenotypes, yet its haplotypic architecture remains incompletely resolved. We refined the genetic landscape of *TMEM106B* by integrating AD-GWAS summary statistics (~978K individuals) with haplotype analyses in 5,873 Dutch genomes including AD-cases, age-matched controls, and cognitively healthy centenarians. In addition to the known linkage disequilibrium (LD) block defining risk and protective haplotypes characterized by the amino acid substitution at residue 185 (Threonine/Serine), we identify a second conditionally independent LD block. Together these form four haplotypes (T1-T4), with T3 strongly enriched in centenarians. Long-read whole-genome sequencing in 493 individuals reveals haplotype-specific

51 methylation patterns and structural variation, including a ~19 Kbp genomic rearrangement carried
52 by T3. These findings indicate that the *TMEM106B* locus comprises multiple haplotypes defined
53 by methylation patterns and structural variants, offering insights into GWAS associations with AD
54 risk and cognitive longevity.
55

INTRODUCTION

TMEM106B plays an important role in maintaining cognitive health during aging. Genome-wide association studies (GWAS) have repeatedly identified variants at the *TMEM106B* locus that influence risk for multiple neurodegenerative diseases [1–3]. These variants are in complete linkage disequilibrium (LD), forming an LD-block that defines two haplotypes. The risk haplotype has a frequency of ~59% in European populations, and has been associated with various neurodegenerative diseases, including Alzheimer's disease [3] and frontotemporal dementia [1,2]. The reciprocal protective haplotype carries the alleles associated with protection against these traits, and occurs at a frequency of approximately ~41%.

We recently observed that the protective haplotype is among the most enriched AD-associated loci in cognitively healthy centenarians [4]. This enrichment suggests that *TMEM106B* may contribute to preserved cognitive function into extreme old age. Importantly, recent neuropathological analyses indicate that *TMEM106B* variants are primarily associated with TDP-43-related pathologies (including LATE-NC and HS) [5]; tau tangles [5,6]; and aggregation of TMEM106B fibrils [7–9]. This raises the possibility that the enrichment of the protective haplotype in centenarians reflects protection against neurodegeneration mediated by these proteinopathies. Consistent with this, a subset of centenarians from the 100-plus cohort exhibited TDP-43 aggregates in the brain, and TDP-43 positivity was associated with reduced cognitive performance [10].

The specific causal variants underlying the effects in the *TMEM106B* haplotypes remain unresolved [11,12]. One candidate is the missense variant rs3173615 (C>T), which results in a Thr185Ser substitution. The risk-associated rs3173615-C allele encodes threonine at position 185 and has been hypothesized to alter glycosylation, potentially influencing protein stability or aggregation propensity [13]. Indeed, carriers of the risk haplotype exhibit higher levels of

82 TMEM106B C-terminal aggregates compared to carriers of the rs3173615-G allele (185-Ser) [7–
83 9]. However, knock-in of the homologous rs3173615-G allele in mice did not confer protection
84 [12], suggesting that this variant alone may not explain disease risk. The risk haplotype contains
85 multiple additional variants in complete LD with rs3173615-C that could contribute to functional
86 differences. These include an AluYb8 insertion in the 3′ untranslated region (3′ UTR) [14–16] and
87 a variant affecting CTCF binding near rs1990620 [17], both of which may influence transcriptional
88 or post-transcriptional regulation of *TMEM106B*.
89
90 Despite consistent GWAS associations, the genetic architecture underlying the *TMEM106B* locus
91 remains incompletely resolved. In particular, variants within the locus are typically inherited in
92 strong linkage disequilibrium, complicating efforts to disentangle independent signals and identify
93 functional alleles. Moreover, structural variants (SVs), including genomic rearrangements,
94 tandem repeats, and transposable elements, are largely invisible to conventional imputation-
95 based approaches and may contribute to haplotype-specific effects that remain unexplored.
96
97 Here, we refine the *TMEM106B* locus by integrating long-read sequencing and large-scale GWAS
98 summary statistics with haplotype analyses in cognitively healthy centenarians, AD cases, and
99 age-matched controls.
100

101 **METHODS**

102 *Study population*
103 Samples used in this study included a total 5,873 genomes: 443 centenarians from the 100-plus
104 Study [18]; 2,321 Alzheimer disease (AD) patients from the Amsterdam Dementia Cohort [19] and
105 the Netherlands Brain Bank [20]; and 3,109 non-demented controls age-matched with the AD
106 patients previously described [4], including 1,614 from LASA cohort [21]. The Medical Ethics

107 Committee of the Amsterdam University Medical Center approved all studies. All participants
108 and/or their legal representatives provided written informed consent for participation in clinical
109 and genetic studies.
110
111 *Genotyping and AD-GWAS summary statistics*
112 Array-based genotyping and haplotype-imputations of the above cohorts were previously
113 described [4].
114
115 We analyzed GWAS variants within 100 Kbp upstream and downstream of *TMEM106B* coding
116 sequence (chr7:12111294-12343367), using the latest AD-GWAS summary statistics [22].
117 Pairwise linkage disequilibrium (LD) values (r^2) were calculated with PLINK (v2.00a6LM) [23]
118 using the centenarian, AD-cases, and control cohorts described above, and variants were
119 clumped using a conservative r^2 threshold of 0.15. We retained the sentinel SNPs of LD blocks
120 with genome-wide significance in the AD-GWAS (p-value < 5.5×10^{-8}). We define an LD block as
121 a set of variants that were in strong LD ($r^2 \geq 0.8$) with the sentinel SNPs, resulting in two blocks:
122 Block 1 (marked by rs5011439) and Block 2 (rs13237415). To avoid overestimation in
123 downstream analysis, we additionally performed pairwise LD between all variants within each LD
124 block, and identified a representative SNP for each block via hierarchical clustering, resulting in
125 two medoid-based marker SNPs for Block 1 (rs12699338) and Block 2 (rs12699361). Joint effect
126 analyses using GCTA-COJO (v1.94.1) [24] were performed separately for (i) sentinel SNPs from
127 each block; and (ii) the corresponding medoid SNPs, confirming that both sets show independent
128 associations.
129
130 *Differential enrichment of TMEM106B haplotype in centenarians*
131 Using the array-imputed haplotypes of the 5,873 genomes (11,746 total haplotypes), we used the

132 sentinel SNPs to classify *TMEM106B* haplotypes (T1-T4), as detailed in Table S1.

133

134 To determine whether individual haplotypes and their genotypes were enriched in centenarians,
135 we used logistic regression models implemented in R (v4.5.2) with the *glm* function by separately
136 comparing centenarians with non-demented controls and AD-cases. All models were adjusted for
137 *APOE* genotype (modeled as categorical factors) and population structure using the first five
138 genetic principal components calculated with PLINK.

139

140 For the enrichment analysis of individual haplotypes, we applied a weighted sum-to-zero (effect)
141 encoding in a joint model to test the independent enrichments of the T1-T4 haplotype. This
142 approach enables estimation of the independent effect of each haplotype relative to the
143 frequency-weighted average effect across all haplotypes, allowing effect sizes (β), standard
144 errors, 95% confidence intervals, and p-values to be interpreted on a common scale. Briefly, we
145 fitted a joint logistic regression model with centenarian phenotype as the response variable,
146 $Centenarian \sim T2 + T3 + T4 + APOE + PC_{1\text{-}5}$, using T1 as the reference level. We then used the
147 estimated β's and their variance-covariance matrix to compute the frequency-weighted average
148 effect across all haplotypes, including T1, where weights were defined by haplotype frequencies
149 in 5,873 array-imputed genomes. We re-expressed the haplotype-specific β's under a weighted
150 sum-to-zero constraint, and applied Wald tests to evaluate whether each haplotype significantly
151 deviated from the weighted average effect.

152

153 To determine specific genotypes of the *TMEM106B* haplotypes, we constructed a categorical
154 variable representing all observed genotypes (e.g., T1/T1, T1/T2, T2/T3, etc.). We then fitted a
155 joint logistic regression model using T1/T1 as the reference genotype, adjusting for *APOE*
156 genotype and genetic $PC_{1\text{-}5}$. Due to the relative low frequency of T4, we adjusted for but not
157 interpreted estimated effects T4-containing genotypes. Odds ratios and 95% confidence intervals

158 were calculated from the regression coefficients.

159

160 To account for multiple testing, we defined two families of tests: (i) enrichment of individual
161 haplotypes and (ii) enrichment of genotypes. We controlled the false discovery rate (FDR) within
162 each family using the Benjamini–Hochberg procedure (family i: 8 tests; family ii: 10 tests). FDR-
163 adjusted q-values <0.1 were considered statistically significant. Adjusted q-values are reported
164 alongside nominal p-values.

165

166 *Long-read sequencing of 493 Dutch genomes*

167 Peripheral blood tissue of a subset of 245 CHCs and 248 ADCs were long-read whole-genome
168 sequenced on Pacific Biosciences (PacBio) Sequel IIe System. For 10 centenarians, we long-
169 read whole-genome sequenced temporal brain tissue. All samples were sequenced at the
170 Department of Clinical Genetics at the Amsterdam University Medical Center (AUMC) and at the
171 Radboud Technology Genome Center and Radboud University Medical Center (RUMC), using 2
172 hours of pre-extension time and 30 hours of collection time. Overall, the AD and centenarian
173 genomes were sequenced at a median coverage of 18.2x, with median read-lengths of 14.8 Kbp.
174 See extended details in **Supplemental Methods**.

175

176 After sequencing, raw reads were collected and analyzed through an in-house pipeline (freely
177 available at https://github.com/holstegelab/snakemake_pipeline; Supplemental Methods).

178

179 In addition, we used publicly available HiFi-based 47 whole-genomes from the Human
180 Pangenome Research Consortium, which were similarly processed through our in-house pipeline
181 (**Supplemental Methods**).

182

183 *Variant and methylation profiles using long-read sequencing data*

184 Single nucleotide polymorphisms (SNPs) were identified in both the ADC and CHC long-read
185 sequencing data using *deepvariant* (v1.1.0) [25], and joint-called using *GLnexus* (v1.4.1-0-
186 g68e25e5) using '--config Deepvariant' parameter [26].

187

188 Structural variants (SVs) in AD cases, centenarians, and HPRC genomes were identified and
189 genotyped using an in-house local assembly pipeline based on Sniffles2 (v2.0.6) [27] and Otter
190 (v1.0) [28] (**Supplemental Methods)**. The pipeline comprises three steps.

191

192 First, we performed local assembly of allele sequences for each candidate SV in each genome.
193 Second, we conducted joint genotyping across genomes using a fast sequence-based clustering
194 procedure implemented in Otter (**Supplemental Results**), enabling direct comparison and
195 harmonization of SV allele sequences across >1 million SV regions and large collections of long-
196 read genomes. This step generated an SV-based gVCF for the centenarian and AD genomes.
197 Third, we recalibrated SV genotypes by probabilistically inferring the most likely genotype based
198 on supporting long-read alignments, together with a corresponding genotype quality (GQ) score,
199 using the *svg* function in Haplr (v0.1.0) (https://github.com/holstegelab/haplr), enabling consistent
200 genotyping across both high-coverage and low-coverage samples. Additionally, this procedure
201 also quantifies CpG methylation within each SV-allele.

202

203 The corresponding allele-sequence of all unique genotypes of all SVs were annotated using
204 DFAM (release 3.7) [29] to annotate transposable elements (TEs), while tandem repeats (TRs)
205 were annotated by overlapping TR-annotation tracks from the UCSC Genome Browser
206 (http://genome.ucsc.edu) [30].

207

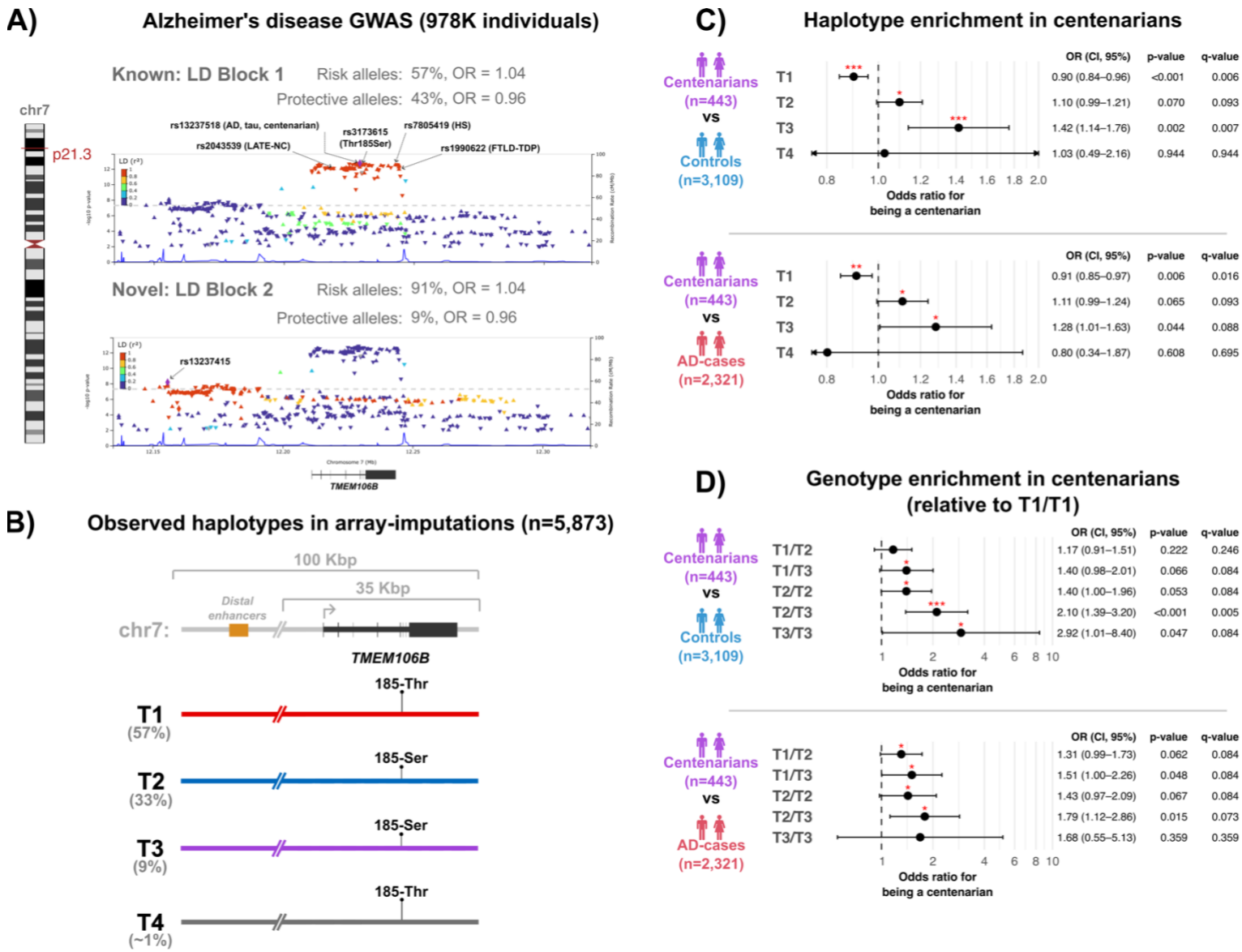
208 To evaluate haplotype-specific structural variants (SVs) and CpG methylation, reads were phased

209 using the *svg* and *mtp* functions in Haplr. GWAS variants defined in (**Table S1**), enabling
210 assignment of SV alleles and CpG methylation to specific haplotypes.
211
212 For methylation analyses, we focused on CpG sites within the start/end coordinates from the
213 union of both LD blocks (chr7:12146976-12247317). CpG sites were retained if they had a
214 minimum per-sample coverage of 4 reads and were observed in at least 10 independent
215 individuals. Methylation levels were calculated per individual and haplotype as the fraction of
216 methylated reads over total reads. CpG methylation was summarised separately for each pairwise
217 haplotype-comparison (T1 vs T2; T1 vs T3; and T2 vs T3) across individuals, and sites with at
218 least 10% median methylation between haplotypes (|Δ methylation| ≥ 0.1) were considered
219 differentially methylated CpG sites, and those with a 90% difference (|Δ methylation| ≥ 0.9) were
220 considered haplotype-specific. Regulatory annotation was incorporated using candidate cis-
221 regulatory element (cCRE) tracks from ENCODE (dataset ENCFF924IMH) [31].
222

223 **RESULTS**

224 *Two independent LD blocks in the TMEM106B locus are associated with Alzheimer's disease*
225 *(AD)*
226
227 To explore whether additional haplotypes, next to the known risk and protective haplotypes also
228 associated with AD-risk, we used linkage disequilibrium (LD) profiles from 5,873 Dutch genomes
229 comprising of cognitively healthy centenarians, AD-cases, and age-matched controls to analyze
230 the summary statistics of the most recent AD-GWAS [22] (~978K participants; **Figure 1A**). We
231 restricted analyses to common variants that (i) surpassed genome-wide significance in their
232 associations with AD ($p \leq 5.5 \times 10^{-8}$), (ii) had minor allele frequency ≥1%, and (iii) clustered within
233 multi-SNP LD blocks defined by $r^2 \geq 0.15$ (see **Methods**).

234


A)
Alzheimer's disease GWAS (978K individuals)
Known: LD Block 1
Risk alleles: 57%, OR = 1.04
Protective alleles: 43%, OR = 0.96
chr7
p21.3
rs13237518 (AD, tau, centenarian)
rs3173615 (Thr185Ser)
rs7805419 (HS)
rs2043539 (LATE-NC)
rs1990622 (FTLD-TDP)
Novel: LD Block 2
Risk alleles: 91%, OR = 1.04
Protective alleles: 9%, OR = 0.96
rs13237415
TMEM106B
B)
Observed haplotypes in array-imputations (n=5,873)
100 Kbp
35 Kbp
Distal enhancers
chr7:
TMEM106B
185-Thr
T1 (57%)
185-Ser
T2 (33%)
185-Ser
T3 (9%)
185-Thr
T4 (~1%)
C)
Haplotype enrichment in centenarians
Centenarians (n=443) vs Controls (n=3,109)
OR (CI, 95%) p-value q-value
T1 0.90 (0.84–0.96) <0.001 0.006
T2 1.10 (0.99–1.21) 0.070 0.093
T3 1.42 (1.14–1.76) 0.002 0.007
T4 1.03 (0.49–2.16) 0.944 0.944
Odds ratio for being a centenarian
Centenarians (n=443) vs AD-cases (n=2,321)
OR (CI, 95%) p-value q-value
T1 0.91 (0.85–0.97) 0.006 0.016
T2 1.11 (0.99–1.24) 0.065 0.093
T3 1.28 (1.01–1.63) 0.044 0.088
T4 0.80 (0.34–1.87) 0.608 0.695
Odds ratio for being a centenarian
D)
Genotype enrichment in centenarians (relative to T1/T1)
Centenarians (n=443) vs Controls (n=3,109)
OR (CI, 95%) p-value q-value
T1/T2 1.17 (0.91–1.51) 0.222 0.246
T1/T3 1.40 (0.98–2.01) 0.066 0.084
T2/T2 1.40 (1.00–1.96) 0.053 0.084
T2/T3 2.10 (1.39–3.20) <0.001 0.005
T3/T3 2.92 (1.01–8.40) 0.047 0.084
Odds ratio for being a centenarian
Centenarians (n=443) vs AD-cases (n=2,321)
OR (CI, 95%) p-value q-value
T1/T2 1.31 (0.99–1.73) 0.062 0.084
T1/T3 1.51 (1.00–2.26) 0.048 0.084
T2/T2 1.43 (0.97–2.09) 0.067 0.084
T2/T3 1.79 (1.12–2.86) 0.015 0.073
T3/T3 1.68 (0.55–5.13) 0.359 0.359
Odds ratio for being a centenarian

235

236 **Figure 1. Four separate *TMEM106B* haplotypes are differentially associated with risk for Alzheimer's disease (AD) and extreme**
237 **cognitive preservation. (A)** LocusZoom plot of the *TMEM106B* locus in the latest AD-GWAS [22] highlighting conditionally
238 independent LD blocks. Top: Block 1 encompassing the known risk and protective haplotypes which harbour alleles previously
239 associated with risk to various neurodegenerative traits and pathologies, marked by rs5011439. Bottom: Block 2 encompassing novel
240 risk and protective haplotypes marked by rs13237415 with similar effects as Block 1. **(B)** The alleles from both blocks are inherited in
241 four specific haplotypes, T1-T4, as observed in 5,873 Dutch genomes. **(C-D)** Enrichment analysis of the haplotypes when comparing
242 centenarians with controls or AD-cases. FDR-adjusted p (10% threshold): *** < 0.01; ** < 0.05; * < 0.1. **(C)** At the individual haplotype
243 level, T3 is significantly enriched in centenarians, while T1 is significantly depleted. **(D)** Relative to T1/T1 homozygous carriers, the
244 T2/T3 genotype is significantly enriched in centenarians.

245 Two independent LD-blocks at *TMEM106B* reached genome-wide significance (**Figure 1A**).
246 Block 1 (marker SNP rs5011439-C; AF=43%, OR=0.96, $p=1.9x10^{-13}$) overlaps the *TMEM106B*
247 coding region and corresponds to known risk and protective haplotypes (risk: rs5011439-G;
248 protective: rs5011439-C). Block 2, marked by SNP rs13237415-T (AF=9%, OR=0.94, $p=7.7x10^{-}$
249 ^{9}), spans ~65 kb upstream to ~25 kb downstream of *TMEM106B* (**Figure 1A**). The blocks are
250 largely independent (r^2=0.11), and each remained significant in joint/conditional GCTA-COJO
251 analyses (OR=0.96 for both; $p=6.4x10^{-9}$ and $3.3x10^{-4}$), even when extending to other marker
252 SNPs in their respective LD blocks (**Methods**). All SNPs in Block 2 are in strong LD (r^2≥0.8; **Table**
253 **S1**), defining two additional haplotypes with rs13237415-G associated with risk and rs13246024-
254 T with protection.
255
256 Together, these findings establish that the *TMEM106B* locus comprises two conditionally
257 independent association signals, each defined by distinct haplotypes.
258
259 *Four TMEM106B haplotypes are differentially enriched in cognitively healthy centenarians*
260
261 The allele combination of the sentinel SNPs of blocks 1 and 2 define four haplotypes across all
262 genomes (n=11,746; **Figure 1B and Table S1**). We refer to these haplotypes as T1-T4 in
263 descending frequency: T1 is the most common combination (57%, n=6,701) and contains the
264 risk-associated alleles of both blocks. T2 follows (33%, n=3,919) and carries the protective alleles
265 of block 1 and risk alleles of block 2. T3 (9%, n=1,020) carries protective alleles from both blocks,
266 and T4 (~1%, n=106) carries risk alleles from block 1 and protective alleles from block 2.
267
268 We found that, compared with non-demented controls, carriers of T3 were significantly associated
269 with a 1.42-fold higher odds of being a centenarian (OR=1.42, 95% CI=[1.14,1.76]; p=0.002;
270 q=0.007). In contrast, T3 was significantly depleted in centenarians, with a 0.9-fold lower odds

271 (OR=0.90, CI=[0.84,0.96]; p=7.43x10^{-4}; q=0.006). T2 had modest significant enrichment
272 (OR=1.10, CI=[0.99,1.21]; p=0.070; q=0.093), while enrichment for T4 was imprecise due to its
273 relative low frequency (OR=1.03, CI=[0.49,2.16]; p=0.944; q=0.944).
274
275 To further disentangle the contribution of the 185-Serine-encoding haplotypes, we repeated the
276 analysis using T2mthe most frequent serine-containing haplotype, as the reference. Under this
277 model, T1 remained depleted (OR=0.82, CI=[0.70,0.96]; p=0.012), while T3 remained enriched
278 (OR=1.29, CI=[1.00,1.66]; p=0.048), indicating that the enrichment of T3 cannot be explained by
279 the serine allele alone and may reflect additional protective features. Similar patterns were
280 observed when comparing centenarians with AD cases, although effect sizes were attenuated
281 **(Figure 1C)**.
282
283 Analogous to the genotype-specific effects observed for ε4/ε4, ε4/ε3, and ε4/ε2 at *APOE*, we next
284 examined the *TMEM106B* genotypes in centenarians. To do this, we fitted a logistic regression
285 model relative to T1/T1 homozygous carriers, the genotype with the most depleted haplotype.
286 This analysis reinforced the prominent role of T3 (**Figure 1D**): the heterozygous T2/T3 genotype
287 had a 2.10-fold significantly higher odds of becoming a centenarian (OR=2.10, CI=[1.39,3.20];
288 p=4.82x10^{-4}; q=0.005), and significantly increased to 2.92 for the homozygous T3/T3 although
289 with wide confidence intervals (OR=2.92, CI=[1.01,8.40]; p=0.047; q=0.084). Other genotypes,
290 including T1/T2, T1/T3, and T2/T2, showed more modest but significant enrichments (**Figure 1D**).
291 In contrast, estimates for genotypes containing T4 were imprecise due to its low frequency.
292 Similar patterns were observed when comparing centenarians with AD cases, although effect
293 sizes were attenuated (**Figure 1D**). All associations were adjusted for *APOE* genotype status.
294
295 *Long-read fine-mapping reveals haplotype-specific structural variants in TMEM106B*
296

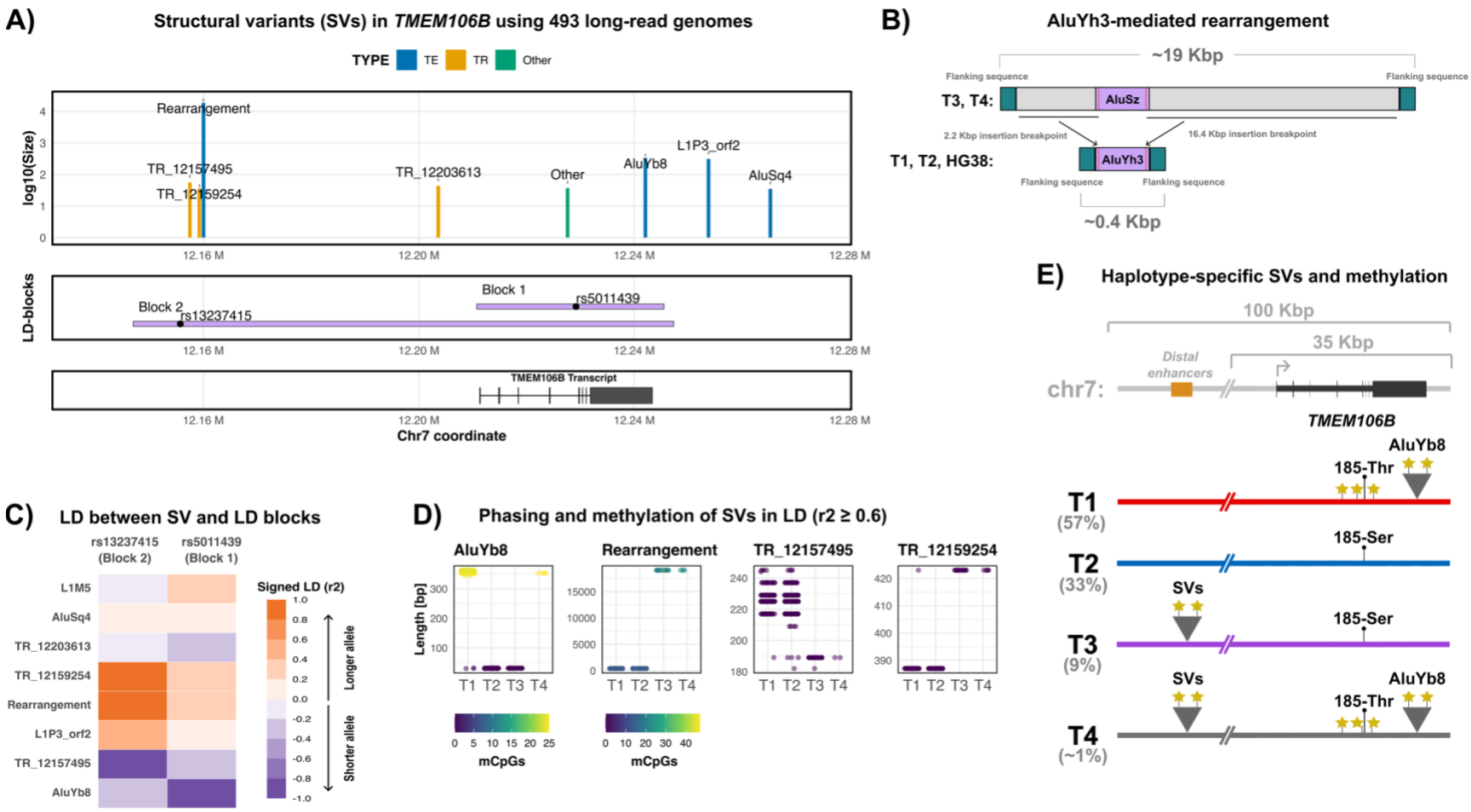
297 To fine-map the genetic basis of the differential associations observed for the T1-T4 *TMEM106B*
298 haplotype, we performed long-read whole-genome sequencing of 493 individuals, including 245
299 AD-cases from the ADC cohort and 248 cognitively healthy centenarians from the 100-plus Study
300 (**Methods**). We used these data to characterize haplotype-resolved structural variation and DNA
301 methylation patterns of the different haplotypes spanning *TMEM106B*.
302
303 We found eight total common structural variants (SVs, ≥25 bp, ≥1% frequency). The SVs ranged
304 from 35 bp to ~19 Kbp (**Figure 2A**), with allele frequencies spanning ~10–90%, similar to those
305 of the two LD blocks (**Table S1**). In addition to the previously reported AluYb8 insertion in the
306 3′UTR, we identified seven novel SVs. The largest was a ~19 Kbp genomic rearrangement located
307 ~51 Kbp upstream of the *TMEM106B* transcription start site, likely mediated by an AluYh3
308 element (**Figure 2B**). Among the identified SVs, a multi-allelic tandem repeat (TR_12157495)
309 located near the rearrangement directly overlapped a known distal enhancer (EH38E2534480),
310 suggesting potential regulatory relevance.
311
312 Both LD-based analyses and long-read phasing independently supported haplotype-specific
313 associations for four SVs (**Figure 2C and 2D**): the AluYb8 insertion was linked to T1 and T4; and
314 the ~19 Kbp rearrangement, along with contractions in TR_12157495 and expansions
315 TR_12159254, were linked to T3 and T4 (**Figure 2C**). The AluYb8 insertion and the ~19 Kbp
316 rearrangement were both highly methylated (**Figure 2D**), highlighting possible regulatory effects.
317
318 We note that while haplotypes T1-T4 are predominantly present in European (Dutch) ancestry,
319 we observed additional haplotypes with distinct allelic combinations in genomes with African
320 ancestry (**Supplemental Results**), consistent with prior work highlighting differences in LD
321 structure in the *TMEM106B* locus between genomes with European and African ancestry [32].
322 This includes a distinct haplotype with the risk-associated 185-Threonine, and the protective-

323 associated absence of the AluYb8 element (**Figure S1**).

324

325 These findings indicate that although both T2 and T3 carry the 185-serine allele, T3 uniquely
326 harbors multiple SVs located ~51 Kbp upstream of *TMEM106B* (**Figure 2E**). These haplotype-
327 specific SVs may contribute to the preferential enrichment of T3 observed in centenarians.
328 Functional studies will be required to determine the biological consequences of these variants.

329



330

331 **Figure 2. The *TMEM106B* haplotypes are characterized by large structural variation (SVs). (A)** Landscape of SVs in the

332 *TMEM106B* locus using 493 long-read sequenced genomes. **(B)** The largest SV was a ~19 Kbp genomic rearrangement likely mediated

333 by an AluYh3 element. **(C)** LD values (r^2) between the SVs and the SNP-markers for the Block 1 and Block 2. **(D)** Long-read phasing

334 and CpG methylation of allele-sequences for SVs in moderate LD ($r^2 \geq 0.6$) with either Block 1 or Block 2. **(E)** These analyses highlight

335 that SVs distinctly characterize the T1-T4 haplotypes of *TMEM106B*.

336 *The TMEM106B haplotypes have distinct methylation profiles*

337

338 We next investigated whether *TMEM106B* haplotypes harbor distinct DNA methylation profiles
339 using CpG methylation frequencies derived from haplotype-resolved long-read sequencing data.
340 Pairwise comparisons of the common haplotypes (T1, T2, and T3) identified 146 unique CpG
341 sites that were differentially methylated in at least one comparison (median methylation difference
342 ≥10%), of which 129 showed near-complete haplotype-specific methylation (median difference
343 ≥90%) (**Figure 3A and Table S2**). Many of these differences were driven by SNPs that created
344 or disrupted CpG dinucleotides, indicating haplotype-specific CpG content (**Table S2**). This
345 included CpGs overlapping known distal enhancers: two linked to T1, three linked to T2, and four
346 linked to T3. Notably, these differentially methylated CpG sites led to higher methylation in the
347 intronic regions (introns 3 and 4) and the 3' UTR in T1, including the AluYb8 element; along with
348 low methylation in T3 in the region overlapping the known rearrangement (**Figure 3A**).

349

350 As the long-read sequencing data were derived from blood, we examined whether similar
351 methylation patterns were observed in temporal cortex tissue by performing long-read sequencing
352 on paired brain samples from 10 centenarians (**Figure 3B**). Across 15 distinct haplotypes, we
353 identified 301 unique CpG sites that were differentially methylated between brain and blood
354 (absolute median methylation difference ≥10%). Tissue-specific patterns were consistent across
355 haplotypes: methylation levels were higher in the brain at the transcription start site, whereas
356 blood showed higher methylation across introns 3 through the 3′UTR. These results indicate that
357 haplotype-associated methylation patterns in *TMEM106B* likely differ by tissues.

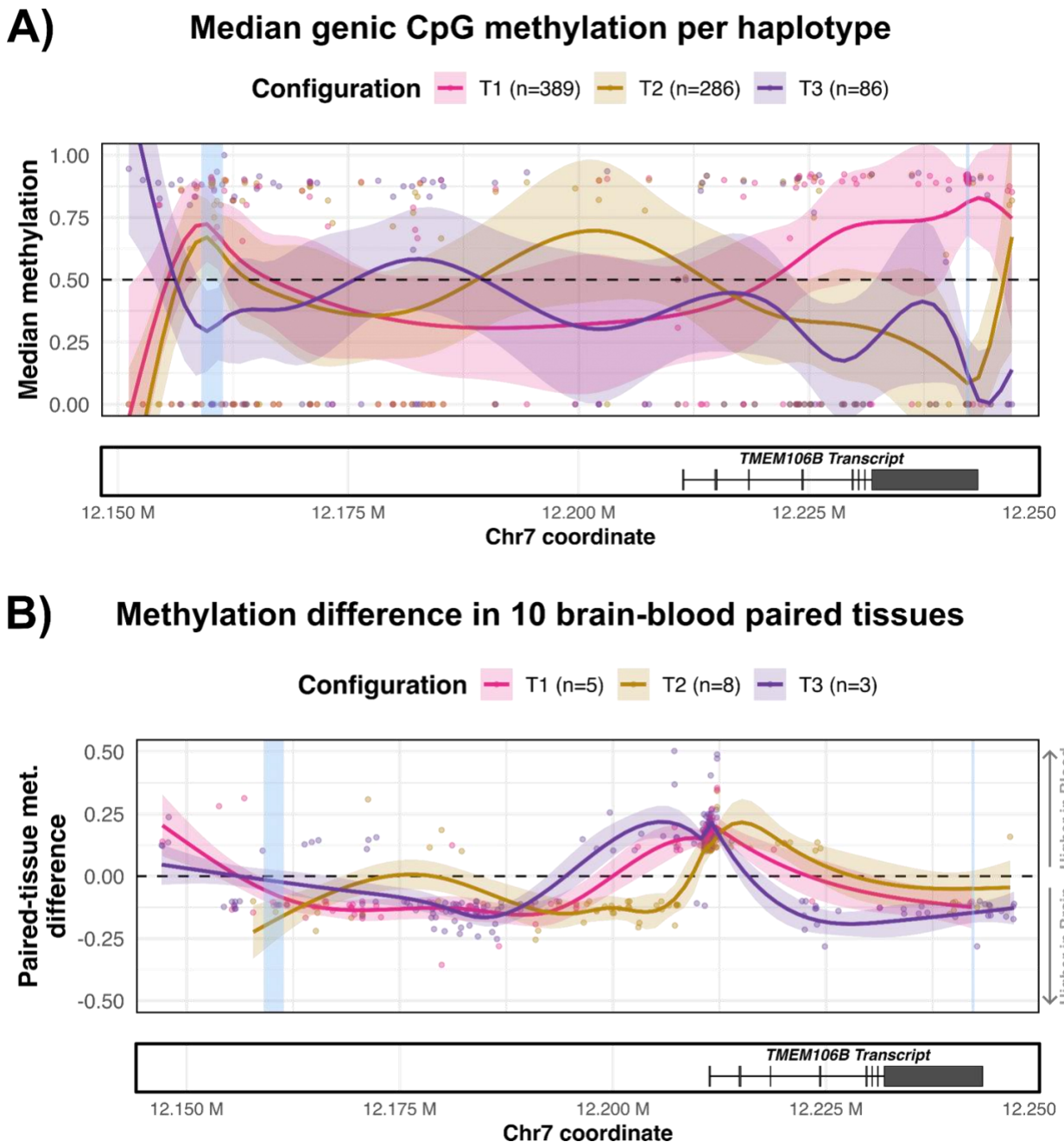


Figure 3. The *TMEM106B* haplotypes are also characterized by distinct CpG DNA methylation profiles. **(A-B)** CpG methylation profiles derived from phased long-read sequencing of blood samples from 493 individuals. Differentially methylated CpG sites are defined as sites where the fraction of methylated reads is at least 10% between haplotypes. **(A)** Using differentially methylated CpG sites (n=146), spline regression of median methylation values per common haplotype (T1-T3) along with standard errors (colour shades), with the bottom plot illustrating the

365 canonical transcript of *TMEM106B*. **(B)** For a subset of 10 individuals, long-read whole-genome
366 sequencing was also performed on matched brain tissue (temporal cortex). Shown are per-
367 sample methylation differences between brain and blood for CpG sites exhibiting an absolute
368 methylation difference ≥10% (n=301), stratified by haplotype. Homozygous haplotype-genotypes
369 were evaluated as a single measurement.

370 **DISCUSSION**

371 In this study, we refine the genetic architecture of the *TMEM106B* locus and demonstrate that two
372 conditionally independent GWAS LD blocks generate four common haplotypes (T1-T4) with
373 differential associations with Alzheimer's disease and extreme cognitive preservation. Among
374 these haplotypes, T3 shows the strongest enrichment in cognitively healthy centenarians, while
375 T1 is depleted. These haplotypes are further distinguished by structural and epigenetic features,
376 including large structural variants (SVs) and distinct DNA methylation patterns. This includes a
377 ~19 Kbp genomic rearrangement ~51 Kbp upstream of the transcription start site of *TMEM106B*
378 that are near and/or overlap known distal enhancers, strongly linked to T3 and T4. In addition, T1
379 is characterized by highly methylated CpG sites within the intronic region and 3′UTR of
380 *TMEM106B*. Together, these findings extend previous GWAS observations and reveal a more
381 nuanced haplotypic structure underlying *TMEM106B*-associated risk.
382
383 Our results do not pinpoint the exact causal variants underlying the associations at the
384 *TMEM106B* locus, but instead indicate that multiple genetic mechanisms likely contribute beyond
385 the previously emphasized amino acid substitution at residue 185 (Threonine/Serine). In
386 particular, the role of the Serine allele and its protective associations with neurodegenerative traits
387 has been actively debated [11,12]. Although both T2 and T3 harbor the serine allele, T3 is more
388 strongly enriched than T2 in centenarians, suggesting that additional protective mechanisms may
389 be encoded within the T3 haplotype. The large structural variants strongly linked to T3 represent
390 plausible candidates for such effects; however, we note that multiple other variants, including CpG
391 sites in regulatory elements, are also in strong linkage disequilibrium with T3 and may contribute
392 to the observed associations.
393
394 *TMEM106B* has been implicated in TMEM106B fibril pathology, TDP-43-related

395 neurodegeneration, and tau pathology [5–9]. The presence of multiple *TMEM106B* haplotypes
396 raises the possibility that they may differentially influence specific pathological pathways, or exert
397 overlapping effects across these processes. The strong enrichment of T3 in cognitively healthy
398 centenarians suggests that this haplotype may contribute to preserving cognitive function [33,34],
399 and potentially confer protection across multiple proteinopathies. Dissecting pathology-specific
400 effects of individual haplotypes will require neuropathologically characterized cohorts and lies
401 beyond the scope of the present study.
402
403 Several limitations warrant consideration. First, our structural variant analysis focused on common
404 SVs (allele frequency ≥1%), and mainly in genomes with European ancestry. Rarer variants, and
405 non-European haplotypes may also contribute to locus heterogeneity [32]. Indeed, we observe
406 additional haplotypes with distinct permutations of the 185-Threonine/Serine amino acids along
407 with the SVs including the AluYb8 element and rearrangement. Evaluating the effects of these
408 novel haplotypes on disease risk is beyond the scope of this study. Second, the T4 haplotype is
409 relatively rare, limiting power to estimate its independent effects with precision. Third, although
410 we identify haplotype-specific structural and methylation differences, functional consequences
411 remain to be experimentally validated. The regulatory impact of the upstream rearrangement,
412 tandem repeats, and AluYb8 elements will require targeted functional assays.
413
414 Despite these limitations, our findings provide new insight into the genetic basis of *TMEM106B*-
415 associated risk and resilience. By leveraging extreme longevity as an informative phenotype and
416 combining haplotype-aware GWAS interpretation with long-read sequencing, we uncover
417 structural and epigenetic features that refine the locus beyond single-marker associations. These
418 results underscore the importance of haplotype-resolved analyses and structural variation in
419 interpreting complex disease loci and suggest that similar hidden architectures may underlie other
420 GWAS signals in neurodegeneration.

ACKNOWLEDGMENTS

The authors are grateful to all study participants, their family members, the participating medical staff, general practitioners, pharmacists and all laboratory personnel involved in patient diagnosis, blood collection, blood biobanking, DNA preparation and sequencing.

SUPPORTING INFORMATION

Supplemental results, methods, figures, and tables can be found in the companion Supplemental Materials and Tables. Consortia members and affiliations can be found in Supplemental Materials.

DATA AND MATERIALS AVAILABILITY

Long-read sequencing data generated with PacBio Sequel 2 for the AD patients cognitively healthy centenarians is available upon submission of a research proposal to the Alzheimer Genetics Hub (AGH, https://alzheimergenetics.org/).

FUNDING

Part of the work in this manuscript was carried out on the Cartesius supercomputer, which is embedded in the Dutch national e-infrastructure with the support of SURF Cooperative. Computing hours were granted to H. H. by the Dutch Research Council (‘100plus’: project# vuh15226, 15318, 17232, and 2020.030; ‘Role of VNTRs in AD’; project# 2022.31, ‘Alzheimer’s Genetics Hub’ project# 2022.38). This work is supported by a VIDI grant from the Dutch Scientific Counsel (#NWO 09150172010083) and a public-private partnership with TU Delft and PacBIo, receiving funding from ZonMW and Health~Holland, Topsector Life Sciences & Health (PPP-allowance), and by Alzheimer Nederland WE.03-2018-07. H.H., S.L., are recipients of ABOARD, a public-private partnership receiving funding from ZonMW (#73305095007) and Health~Holland, Topsector Life Sciences & Health (PPP-allowance; #LSHM20106). S.L. is recipient of ZonMW funding (#733050512). H.H. was supported by the Hans und Ilse Breuer Stiftung (2020),

Dioraphte 16020404 (2014) and the HorstingStuit Foundation (2018). Acquisition of the PacBio Sequel II long read sequencing machine was supported by the ADORE Foundation (2022).

CONSENT OF STATEMENT

The Medical Ethics Committee of the Amsterdam University Medical Center approved all studies. All participants and/or their legal representatives provided written informed consent for participation in clinical and genetic studies.

CONFLICT OF INTEREST STATEMENT

HH has a collaboration contract with Muna Therapeutics, PacBio, Neurimmune and Alchemab. She serves in the scientific advisory boards of Muna Therapeutics and is an external advisor for Retromer Therapeutics. Research of Alzheimer center Amsterdam is part of the neurodegeneration research program of Amsterdam Neuroscience. Alzheimer Center Amsterdam is supported by Stichting Alzheimer Nederland and Stichting Steun Alzheimercentrum Amsterdam. The clinical database structure was developed with funding from Stichting Dioraphte.

462 **REFERENCES**

463 [1] Van Deerlin VM, Sleiman PMA, Martinez-Lage M, Chen-Plotkin A, Wang L-S, Graff-
464 Radford NR, et al. Common variants at 7p21 are associated with frontotemporal lobar
465 degeneration with TDP-43 inclusions. Nat Genet 2010;42:234–9.
466 https://doi.org/10.1038/ng.536.
467 [2] Pottier C, Küçükali F, Baker M, Batzler A, Jenkins GD, Van Blitterswijk M, et al.
468 Deciphering distinct genetic risk factors for FTLD-TDP pathological subtypes via whole-
469 genome sequencing. Nat Commun 2025;16:3914. https://doi.org/10.1038/s41467-025-
470 59216-0.
471 [3] Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, et al. New
472 insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet
473 2022;54:412–36. https://doi.org/10.1038/s41588-022-01024-z.
474 [4] Tesi N, Van Der Lee S, Hulsman M, Van Schoor NM, Huisman M, Pijnenburg Y, et al.
475 Cognitively healthy centenarians are genetically protected against Alzheimer's disease.
476 Alzheimer's & Dementia 2024;20:3864–75. https://doi.org/10.1002/alz.13810.
477 [5] Shade LMP, Katsumata Y, Abner EL, Aung KZ, Claas SA, Qiao Q, et al. GWAS of multiple
478 neuropathology endophenotypes identifies new risk loci and provides insights into the
479 genetic risk of dementia. Nat Genet 2024;56:2407–21. https://doi.org/10.1038/s41588-024-
480 01939-9.
481 [6] Vialle RA, De Paiva Lopes K, Li Y, Ng B, Schneider JA, Buchman AS, et al. Structural
482 variants linked to Alzheimer's disease and other common age-related clinical and
483 neuropathologic traits. Genome Med 2025;17:20. https://doi.org/10.1186/s13073-025-
484 01444-6.

485 [7] T. Vicente C, Perneel J, Wynants S, Heeman B, Van Den Broeck M, Baker M, et al. C-
486 terminal TMEM106B fragments in human brain correlate with disease-associated
487 *TMEM106B* haplotypes. Brain 2023;146:4055–64. https://doi.org/10.1093/brain/awad133.
488 [8] Lee JY, Harney DJ, Teo JD, Kwok JB, Sutherland GT, Larance M, et al. The major
489 TMEM106B dementia risk allele affects TMEM106B protein levels, fibril formation, and
490 myelin lipid homeostasis in the ageing human hippocampus. Mol Neurodegeneration
491 2023;18:63. https://doi.org/10.1186/s13024-023-00650-3.
492 [9] Marks JD, Ayuso VE, Carlomagno Y, Yue M, Todd TW, Hao Y, et al. TMEM106B core
493 deposition associates with TDP-43 pathology and is increased in risk SNP carriers for
494 frontotemporal dementia. Sci Transl Med 2024;16:eadf9735.
495 https://doi.org/10.1126/scitranslmed.adf9735.
496 [10] Zhang M, Ganz AB, Rohde S, Lorenz L, Rozemuller AJM, Van Vliet K, et al. The
497 correlation between neuropathology levels and cognitive performance in centenarians.
498 Alzheimer's & Dementia 2023;19:5036–47. https://doi.org/10.1002/alz.13087.
499 [11] Jiao H-S, Yuan P, Yu J-T. TMEM106B aggregation in neurodegenerative diseases: linking
500 genetics to function. Mol Neurodegeneration 2023;18:54. https://doi.org/10.1186/s13024-
501 023-00644-1.
502 [12] Cabron A-S, Borgmeyer U, Richter J, Peisker H, Gutbrod K, Dörmann P, et al. Lack of a
503 protective effect of the Tmem106b "protective SNP" in the Grn knockout mouse model for
504 frontotemporal lobar degeneration. Acta Neuropathol Commun 2023;11:21.
505 https://doi.org/10.1186/s40478-023-01510-3.
506 [13] Nicholson AM, Finch NA, Wojtas A, Baker MC, Perkerson RB, Castanedes-Casey M, et al.
507 TMEM 106B p.T185S regulates TMEM 106B protein levels: implications for frontotemporal
508 dementia. Journal of Neurochemistry 2013;126:781–91. https://doi.org/10.1111/jnc.12329.
509 [14] Chemparathy A, Le Guen Y, Zeng Y, Gorzynski J, Jensen TD, Yang C, et al. A 3′UTR
510 Insertion Is a Candidate Causal Variant at the *TMEM106B* Locus Associated With

511 Increased Risk for FTLD-TDP. Neurol Genet 2024;10:e200124.
512 https://doi.org/10.1212/NXG.0000000000200124.
513 [15] Rodney A, Karanjeet K, Benzow K, Koob MD. A common Alu insertion in the 3'UTR of
514 *TMEM106B* is associated with risk of dementia. Alzheimer's & Dementia
515 2024;20:5071–7. https://doi.org/10.1002/alz.14090.
516 [16] Salazar A, Tesi N, Knoop L, Pijnenburg Y, Van Der Lee S, Wijesekera S, et al. An AluYb8
517 retrotransposon characterises a risk haplotype of TMEM106B associated in
518 neurodegeneration 2023. https://doi.org/10.1101/2023.07.16.23292721.
519 [17] Gallagher MD, Posavi M, Huang P, Unger TL, Berlyand Y, Gruenewald AL, et al. A
520 Dementia-Associated Risk Variant near TMEM106B Alters Chromatin Architecture and
521 Gene Expression. The American Journal of Human Genetics 2017;101:643–63.
522 https://doi.org/10.1016/j.ajhg.2017.09.004.
523 [18] Holstege H, Beker N, Dijkstra T, Pieterse K, Wemmenhove E, Schouten K, et al. The 100-
524 plus Study of cognitively healthy centenarians: rationale, design and cohort description.
525 Eur J Epidemiol 2018;33:1229–49. https://doi.org/10.1007/s10654-018-0451-3.
526 [19] Van Der Flier WM, Scheltens P. Amsterdam Dementia Cohort: Performing Research to
527 Optimize Care. Journal of Alzheimer's Disease 2018;62:1091–111.
528 https://doi.org/10.3233/JAD-170850.
529 [20] The Netherlands Brain Bank for Psychiatry. Handbook of Clinical Neurology, vol. 150,
530 Elsevier; 2018, p. 3–16. https://doi.org/10.1016/B978-0-444-63639-3.00001-3.
531 [21] Hoogendijk EO, Van Schoor NM, Qi Y, Visser M, Swinkels JC, Broese Van Groenou MI, et
532 al. The Longitudinal Aging Study Amsterdam: design and cohort update 2025. Eur J
533 Epidemiol 2025;40:705–20. https://doi.org/10.1007/s10654-025-01238-5.
534 [22] ADGC, Bonn, CHARGE, EADB, Bellenguez C, EADI, et al. Consensus meta-analysis of
535 genome-wide association studies for Alzheimer's disease and related dementia 2025.
536 https://doi.org/10.1101/2025.10.20.25338060.

537 [23] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A
538 Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The
539 American Journal of Human Genetics 2007;81:559–75. https://doi.org/10.1086/519795.
540 [24] Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex
541 Trait Analysis. The American Journal of Human Genetics 2011;88:76–82.
542 https://doi.org/10.1016/j.ajhg.2010.11.011.
543 [25] Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP
544 and small-indel variant caller using deep neural networks. Nat Biotechnol 2018;36:983–7.
545 https://doi.org/10.1038/nbt.4235.
546 [26] Yun T, Li H, Chang P-C, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant
547 calls using DeepVariant and GLnexus. Bioinformatics 2021;36:5582–9.
548 https://doi.org/10.1093/bioinformatics/btaa1081.
549 [27] Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al.
550 Comprehensive Structural Variant Detection: From Mosaic to Population-Level 2022.
551 https://doi.org/10.1101/2022.04.04.487055.
552 [28] Tesi N, Salazar A, Zhang Y, Van Der Lee S, Hulsman M, Knoop L, et al. Characterizing
553 tandem repeat complexities across long-read sequencing platforms with TREAT and *otter*.
554 Genome Res 2024;34:1942–53. https://doi.org/10.1101/gr.279351.124.
555 [29] Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of
556 transposable element families, sequence models, and genome annotations. Mobile DNA
557 2021;12:2. https://doi.org/10.1186/s13100-020-00230-y.
558 [30] Casper J, Speir ML, Raney BJ, Perez G, Nassar LR, Lee CM, et al. The UCSC Genome
559 Browser database: 2026 update. Nucleic Acids Research 2026;54:D1331–5.
560 https://doi.org/10.1093/nar/gkaf1250.

561 [31] The ENCODE Project Consortium, Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V,
562 et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes.
563 Nature 2020;583:699–710. https://doi.org/10.1038/s41586-020-2493-4.
564 [32] Katsumata Y, Fardo DW, Shade LMP, Alzheimer's Disease Genetics Consortium, Bowen
565 JD, Crane PK, et al. LATE-NC risk alleles (in *TMEM106B, GRN* , and *ABCC9* genes)
566 among persons with African ancestry. Journal of Neuropathology & Experimental
567 Neurology 2023;82:760–8. https://doi.org/10.1093/jnen/nlad059.
568 [33] Rohde SK, Luimes MC, Lorenz LMC, Fierro-Hernández P, Rozemuller AJM, Hulsman M,
569 et al. Amyloid-Beta Pathology and Cognitive Performance in Centenarians. JAMA Neurol
570 2025;82:837. https://doi.org/10.1001/jamaneurol.2025.1734.
571 [34] Zhang M, Ganz AB, Rohde S, Rozemuller AJM, Bank NB, Reinders MJT, et al. Resilience
572 and resistance to the accumulation of amyloid plaques and neurofibrillary tangles in
573 centenarians: An age-continuous perspective. Alzheimer's & Dementia
574 2023;19:2831–41. https://doi.org/10.1002/alz.12899.
575