# FINE-MAPPING OF THE TMEM106B LOCUS REVEALS FOUR HAPLOTYPES THAT ARE DIFFERENTIALLY ASSOCIATED WITH RISK FOR NEURODEGENERATION

**AUTHORS:**
Alex Salazar[1], Niccolò Tesi[1,2], Lydian Knoop[1], Yolande Pijnenburg[3,4], Sven van der Lee[1,4], Sanduni Wijesekera[1], Jana Krizova[1], Mikko Hiltunen[5], Markus Damme[6], Leonard Petrucelli[7,8], Marcel Reinders[2], August B. Smit[9], Marc Hulsman[1], ADGC, Bonn, CHARGE, EADB, EADI, FinnGen, GERAD, GR@ACE/DEGESCO, PGC-ALZ, and Henne Holstege[1,3,4]

**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. Alzheimer Center Amsterdam, Neurology, Vrije Universiteit Amsterdam, Amsterdam UMC location VUmc, Amsterdam, The Netherlands
4. Amsterdam Neuroscience, Neurodegeneration, Amsterdam, The Netherlands
5. Institute of Biomedicine, University of Eastern Finland. Kuopio, Finland
6. Institute of Biochemistry, Christian-Albrechts-University Kiel, Otto-Hahn-Platz 9, Kiel, Germany
7. Department of Neuroscience, Mayo Clinic, Jacksonville, FL, USA
8. Neuroscience Graduate Program, Mayo Clinic Graduate School of Biomedical Sciences, Jacksonville, FL, USA
9. Department of Molecular and Cellular Neurobiology, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

**Corresponding author:**
Henne Holstege: h.holstege@amsterdamumc.nl

**ONE SENTENCE SUMMARY**
GWAS and long-read sequencing suggests that the neurodegenerative genetic etiology associated with *TMEM106B* is conferred by multiple independent haplotypes.

**ABSTRACT**
Genome-wide association studies (GWAS) linked *TMEM106B* variants to susceptibility for neurodegenerative diseases, but the causal genetic elements remain unclear.

Genotyping of 5,792 Alzheimer disease cases and controls revealed that the *TMEM106B* locus consists of four major haplotypes: HA/Ha (covering the coding region), and HB/Hb (covering the upstream regulatory region). These combine into four combinations with varying population-frequencies: *HAB* (57%), *HaB* (34%), *Hab* (9%), and *HAb* (<1%).

Long-read sequencing of 513 individuals showed that *HA* haplotypes (marked by 185-Threonine) carry unique methylated CpG sites and an AluYb8-retrotransposon in the 3' UTR, while the *Ha* haplotypes are marked by the 185-Serine allele. *Hb* haplotypes carry several structural variants (SVs) in nearby distal enhancers, including a 19 Kbp rearrangement, absent in all other haplotypes.

Joint association models revealed that the *HAB* combination (AluYb8+185-Threonine) is risk-increasing, while *Hab* (SVs+185-Serine) confers the protective effect. *HaB* (185-Serine only) is neutral, while *HAb* was too rare to assess. Indeed, relative to middle-aged non-demented controls, cognitively healthy centenarians were more enriched with *Hab* (OR=1.49, $p_{adj}=2.18\times10^{-2}$) than with *HaB* (OR=1.23, $p_{adj}=5.06\times10^{-2}$). Finally, proteomic analysis of temporal cortex tissues (n=182) indicated that relative to the neutral *HaB* combination, the protective *Hab* is associated with 0.91-fold lower TMEM106B C-terminal peptide abundance (CI=[0.81-1.02]), while the risk-increasing *HAB* is associated with 1.07-fold higher abundance (CI=[0.99-1.15]).

Together, this presents a complex genetic structure underlying the association of the *TMEM106B* locus with neurodegenerative diseases.

## INTRODUCTION

*TMEM106B* is a gene that plays a central role in maintaining cognitive health with increasing age. Previous genome-wide association studies (GWAS) have identified genetic variants in *TMEM106B* that are associated with increased risk for various neurodegenerative diseases. This includes the rs13237518 SNP, associated with Alzheimer's disease (AD) *(1)*; rs1990622, associated with frontotemporal lobar degeneration with TDP-43 (FTLD-TDP) *(2)*; rs2043539 associated with limbic-predominant age-related TDP-43 encephalopathy neuropathological change (LATE-NC) *(3)*; and rs7805419 associated with hippocampal sclerosis (HS) *(3)*. These GWAS variants are in complete linkage disequilibrium (LD), and hence, inherited together in the same haplotype **(Fig. 1A)**. As such, the genetic susceptibility to various neurodegenerative diseases as identified by GWAS studies is united in two common haplotypes, which we refer to as *HA* and its reciprocal *Ha*. Haplotype *HA* carries the alleles associated with increased risk for neurodegeneration, and has a frequency of ~60% in Europeans. While Ha carries the reciprocal protective alleles, with a frequency of ~40%.

Recently, we found that of all AD-associated GWAS loci, the alleles in the protective *TMEM106B Ha* haplotype are among the most enriched in a cohort of cognitively healthy centenarians *(4)*. This suggests that *Ha* plays a crucial role in maintaining cognitive health until extreme age. Notably, recent evidence suggests that *TMEM106B* associates specifically with TDP43-related limbic-predominant age-related TDP-43 encephalopathy (LATE) neuropathologic change (LATE-NC) and hippocampal sclerosis (HS), and not with AD-related amyloid-beta and tau neuropathology *(3)*; suggesting that the enrichment in centenarians specifically prevented the development of LATE. Indeed, we previously showed that a small subset of the centenarians from the 100-plus cohort were positive for TDP-43 aggregates, and that TDP-43 positivity associated with a lower cognitive performance *(5)*. Hence, the associated loci reported in the Bellenguez *et al.* GWAS likely linked to broader dementia endophenotypes.

The *TMEM106B HA* risk haplotype was previously observed to be associated with increased TMEM106B protein aggregation into amyloid fibrils in neurodegenerative and aged human brains *(6–8)*. The genetic variants on *HA* that drive TMEM106B aggregation are unclear *(9, 10)*, but multiple candidate-variants have been proposed. One of these candidates is the missense variant, rs3173615 C>T, which encodes a Thr185Ser substitution. The risk-increasing rs3173615-C allele encodes a threonine at residue 185, which is hypothesized to promote differential glycosylation *(11)*, potentially promoting downstream brain aggregation *(9)*. Indeed, individuals that carry the HA haplotype are more likely to have increased levels of TMEM106B C-terminal aggregates than carriers of the rs3173615-G allele (185-Ser) *(6–8)*. However, a knock-in experiment of the homologous rs3173615-G allele in mice showed no protective effect for this allele *(10)*, suggesting a neutral mechanistic role. Indeed, the *HA* haplotype encompasses several other variants that are in complete linkage disequilibrium (LD) with the rs3173615-C risk allele, which may also contribute to disease risk. This includes an AluYb8 element inserted in the 3' UTR of the *TMEM106B* gene *(12–14)*, and a CTCF-modifying variant close to rs1990620 *(15)*; both of which may influence regulation of the *TMEM106B* transcription and translation.

In this study, we set out to fine-map *TMEM106B* locus to better characterize genetic variants that may drive allelic differences in neurodegenerative risk. In contrast to previous studies, we focused on large structural variants (SVs) such as genomic rearrangements, variable tandem repeats (TRs) and transposable elements (TEs). Although these types of SVs can regulate nearby genes and contribute to neurological disorders *(16)*, they have been historically difficult to characterize *(17, 18)*. Given their exclusion in GWAS and existing fine-mapping approaches suggests that the catalog of genetic variation and haplotypes that characterize GWAS loci, including *TMEM106B,*

remains incomplete *(10, 11)*. Hence, we first used genotyping data of cognitively healthy centenarians (CHCs), AD cases (ADCs), and AD-age-matched non-demented individuals to identify *TMEM106B* haplotypes that independently contribute to susceptibility for neurodegeneration. We then characterized SVs and methylation patterns in the haplotypes by using long-read whole-genome sequencing data for a subset of CHCs and ADCs genomes. Through the use of CHCs as 'extreme' control, we assessed how diverse *TMEM106B* haplotypes are differentially associated with neurodegeneration.

## RESULTS

### Four haplotypes in the *TMEM106B* locus influence genetic susceptibility to neurodegeneration

We first aimed to identify additional haplotypes within the *TMEM106B* locus that are associated with AD, independent from the *HA* and *Ha* haplotypes. To do this, we used the latest AD-GWAS summary statistics from Bellenguez *et al.* 2023 *(19)*, along with LD-patterns of imputed genotype data from 346 cognitively healthy centenarians *(4, 20)*, 2,281 AD cases *(21)* (referred to as ADCs), and 3,165 non-demented individuals age-matched to ADCs *(4)* (referred to as NDCs; see **Methods**). We focused on *common variants* that are genome-wide significant AD-GWAS (p-value $< 5x10^{-8}$) and allele frequency (AF) ≥ 2.5%, using a $r^2$ threshold of 0.15 **(Fig. 1A;** see Methods**).**

We identified two separate LD-blocks with the corresponding representative GWAS variants: rs4721057-G (AF = 42%) and rs13246024-T (AF = 9.6%) (**Fig. 1B** and **1C;** see **Supplemental Results)**. A joint effect analysis with *GCTA-COJO (22)* showed that both variants were independently associated with AD in the Bellenguez *et al.* GWAS (ORs of 0.964 and 0.959, respectively; FDR-adjusted q-values of $2.07x10^{-7}$ and $3.27x10^{-4}$, respectively) (**Fig. 1D and Table S1)**.

Of the two variants, rs4721057 associated strongest with AD in the Bellenguez *et al*. GWAS (p-value = $1.14x10^{-11}$), and it is in complete LD with the rs3173615 coding variant ($r^2$ = 0.99). Hence, the rs4721057-G allele is part of the well-known *TMEM106B HA* risk haplotype, and the reciprocal rs4721057-A allele is part of the reciprocal *Ha* protective haplotype **(Table 2** and **Fig. 1E)**.

The second variant, rs13246024, is also associated with AD in the Bellenguez *et al.* GWAS (p-value = $3.02x10^{-8}$), with a similar effect size as the known *HA* and *Ha* haplotypes, and is in low LD with both rs4721057 ($r^2$ = 0.11) and rs3173615 ($r^2$ = 0.11). A closer inspection of the *TMEM106B* locus revealed that rs13246024 is in strong LD ($r^2$ ≥ 0.95) with 27 variants, all located 21-64 Kbps upstream of the *TMEM106B* transcription start site **(Fig. 1C)**. This suggests that rs13246024 defines two separate *TMEM106B* haplotypes **(Fig. 1E)**, which we refer to as *HB* and *Hb* (**Table 2** and **Fig. 1E**).

**Table 1. Samples used in this study.** The genomes of CHCs, ADCs and NDCs (**Methods**) were genotyped using either global screening array (GSA) and/or PacBio HiFi long-read whole-genome sequencing. Proteomic profiles were also analysed for a subset of individuals. Shown are the total number of genomes and proteomic profiles generated for each cohort and tissue across each technology. For 10 CHCs, long-read genomes were generated for one of their offspring.

| Cohort | Tissue | Array imputed genotypes | Long-read sequenced genomes | Mass spectrometry proteomics |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Cognitively healthy centenarians (CHCs) | Peripheral blood | 346 | 245 | |
| | Temporal cortex | | 10 | 55 |
| Offspring of CHCs | Peripheral blood | | 10 | |
| Alzheimer's disease patients (ADCs) | Peripheral blood | 2281 | 248 | |
| | Temporal cortex | | | 85 |
| Non-demented individuals (NDCs) | Peripheral blood | 3165 | | |
| | Temporal cortex | | | 42 |

In agreement with previous studies, CHCs were significantly enriched with *Ha* when compared to NDCs (marginal model OR = 1.28, CI = [1.09-1.50]; p-value = $9.43 \times 10^{-3}$; FDR-adjusted) **(Table S3).** However, we found that the *Hb* haplotype was also significantly enriched in CHCs when compared to NDCs (marginal model OR = 1.33, 95% confidence interval [CI] = [1.031-1.69]; p-value = $3.58 \times 10^{-2}$; FDR-adjusted). This suggests that the *Hb* haplotype may confer additional protection against neurodegeneration in CHCs. Notably, the enrichment of both *Ha* and *Hb* in CHCs is stronger when comparing with NDCs instead of ADCs **(Table S3).**

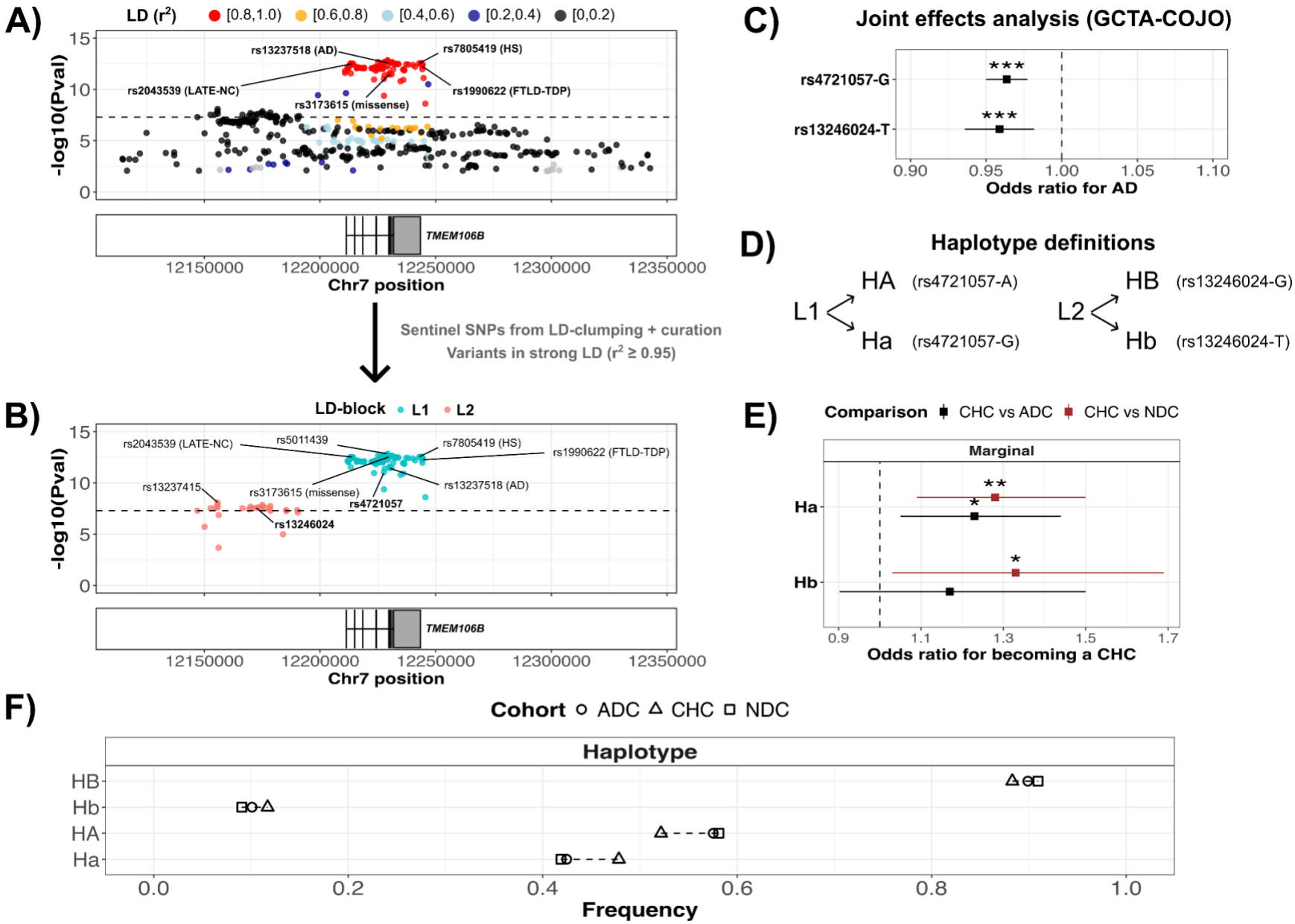
We hence set out to fine-map the *TMEM106B* haplotypes, by investigating the presence of large structural variants, as these may have differential regulatory effects on *TMEM106B* expression and function.

**Table 2. *TMEM106B* haplotype definitions based on the genotypes of imputable GWAS variants.** The marker variants, rs4721057 and rs13246024, are based on linkage disequilibrium (LD) patterns within the CHC, ADC, and NDC cohorts (see **Results and Methods**). While the marker variants, rs5011439 and rs13237415, are based on the sentinel SNPs from clumping analysis using Stage 1 GWAS summary statistics from Bellenguez *et al.* 2023 *(19)*.

| Haplotype | Alleles of marker variants in CHC, ADC, and NDC cohorts | Sentinel alleles in AD-GWAS | Notable variants |
|---|---|---|---|
| **HA** | rs4721057-A | rs5011439-G | rs3173615-C (Threonine); AluYb8 element |
| **Ha** | rs4721057-G | rs5011439-C | rs3173615-G (Serine) |
| **HB** | rs13246024-C | rs13237415-T | |
| **Hb** | rs13246024-G | rs13237415-G | 19 Kbp genome rearrangement; shortest alleles $(AAGG)_n$ of tandem |

| | | | |
|---|---|---|---|
| | | | repeat ($TR_{12157476}$) |

181


A)
LD (r²) [0.8,1.0) [0.6,0.8) [0.4,0.6) [0.2,0.4) [0,0.2)
-log10(Pval)
rs13237518 (AD)
rs7805419 (HS)
rs2043539 (LATE-NC)
rs1990622 (FTLD-TDP)
rs3173615 (missense)
TMEM106B
12150000 12200000 12250000 12300000 12350000
Chr7 position
Sentinel SNPs from LD-clumping + curation
Variants in strong LD (r² ≥ 0.95)
B)
LD-block L1 L2
rs2043539 (LATE-NC)
rs5011439
rs7805419 (HS)
rs1990622 (FTLD-TDP)
rs13237415
rs3173615 (missense)
rs13237518 (AD)
rs4721057
rs13246024
TMEM106B
Chr7 position
C)
Joint effects analysis (GCTA-COJO)
rs4721057-G
rs13246024-T
0.90 0.95 1.00 1.05 1.10
Odds ratio for AD
D)
Haplotype definitions
L1
HA (rs4721057-A)
Ha (rs4721057-G)
L2
HB (rs13246024-G)
Hb (rs13246024-T)
E)
Comparison CHC vs ADC CHC vs NDC
Marginal
Ha
Hb
0.9 1.1 1.3 1.5 1.7
Odds ratio for becoming a CHC
F)
Cohort ADC CHC NDC
Haplotype
HB
Hb
HA
Ha
0.0 0.2 0.4 0.6 0.8 1.0
Frequency

182

**Figure 1**. **The *TMEM106B* locus harbors four haplotypes that independently contribute to neurodegenerative risk. (A-B)** GWAS variants in the *TMEM106B* locus using Stage 1 GWAS summary statistics from Bellenguez et al. 2023 *(19)*. Y-axis: p-values of association with AD; X-axis: chromosome 7 location. Variants are shown in reference to rs3173615 using LD values ($r^2$) calculated from 346 cognitively healthy centenarians (CHCs), 2,281 AD cases (ADCs), and 3,165 non-demented controls (NDCs) (see Methods). Only common variants with allele frequency ≥ 2.5% are shown. **(A)** Top: GWAS variants with p-values < 0.01; dashed line indicates genome-wide significance ($p < 5x10^{-8}$). Key neurodegenerative-associated variants are labeled: rs13237518 (AD), rs1990622 (FTLD-TDP), rs2043539 (LATE-NC), and rs7805419 (HS). Bottom: TMEM106B canonical transcript (ENST00000396668.8); exons and 3' UTR shown as grey boxes. **(B)** Sentinel SNPs rs5011439 and rs13237415 identified by clumping with PLINK ($r^2$ threshold of 0.15; p < 0.001). Variants in complete LD ($r^2 \geq 0.95$) with rs5011439 (blue; n = 106) or rs13237415 (red; n = 27). Separate LD clustering analysis revealed rs4721057 and rs13246024 as marker variants for L1 and L2 within CHC, NDC, and ADC cohorts. **(C)** GCTA-COJO joint effects analysis shows independent association of L1 and L2 LD-blocks with AD. FDR-adjusted significant associations (q ≤ 0.001) of the GCTA-COJO-adjusted p-values are marked by three asterisks. **(D)** Haplotype definitions for L1 and L2: risk-increasing HA haplotype carries rs3173615-C (185-Thr); the protective Ha haplotype carries rs3173615-G (185-Ser). **(G)** Phased haplotype imputations reveal four unique haplotype combinations. **(H)** Enrichment analysis (marginal logistic regression model) of *Ha* and *Hb* in CHCs relative to ADCs (black) and NDCs (brown) in similar layout as Fig. 1C, where one and two asterisks indicate q-value ≤ 0.05 and q-value ≤ 0.001, respectively.

***TMEM106B* haplotypes carry unique combinations of SVs affecting ~20 Kbp total genomic sequence**

We used long-read whole-genome sequencing of 493 genomes derived from 245 AD patients from the ADC cohort and 248 CHCs from the 100-plus Study cohort (**Methods** and **Supplemental Results**), to generate a comprehensive catalog of large structural variants (SVs) in the *TMEM106B* locus. We developed a long-read SV-calling procedure to characterize SVs caused by variable tandem repeats (TRs), transposable elements (TEs), and large genomic rearrangements with high-resolution (**Methods** and **Supplemental Results**). Focusing on common variation (AF ≥ 2.5%) in a 100 Kbp window upstream and downstream of the *TMEM106B* coding sequence (232 Kbp total), we identified 18 SVs (> 15 bp): 7 caused by variable TEs, and 11 caused by variable TRs (**Fig. 2A, Table S6,** and **Supplemental Results**).

An in-depth investigation of the 18 common SVs based on existing annotations of regulatory elements, as well as sequence similarity to SVs with known associations with human diseases, revealed that six SVs may influence transcription and/or translation of *TMEM106B* (detailed in **Supplemental Results**); all located in a 109 Kbp region encompassing the *TMEM106B* coding region (chr7:12146000-12255000). We characterize the six SVs below.

In agreement with previous studies, the AluYb8 element (AF = 56%) was significantly enriched in the *HA* risk haplotype (FDR-adjusted q-value = $1x10^{-308}$, see **Methods) (Fig. 2B)**. Additionally, we identified a bi-allelic TR located in intron 4, referred to as $TR_{1222762}$, where the expanded allele (≥ 60 bp) was enriched with the *Ha* protective haplotype (q-value = $8.45x10^{-15}$).

Three SVs in a distal-enhancer cluster ~51 Kbp upstream of the *TMEM106B* transcription start site distinctly characterise the *Hb* haplotype (**Fig. 2C**). Two rearrangement-based SVs, *RA_1* and *RA_2*, affect 2.2 and 16.5 Kbp of genomic sequence, respectively **(Fig. 2D)**, and occur together (AF = ~11%) in complete LD with each other (see next section). They are more enriched in the *Hb* haplotype (q-values of ~$2.30x10^{-132}$) than in the *Ha* haplotype (q-values of $6.59x10^{-10}$ and $6.73x10^{-10}$, respectively). We speculate that they may have arisen from a single rearrangement-event mediated by a human-specific insertion of an AluYh3 TE **(Fig. 2D and Supplemental Results**); similar to a rearrangement in the *PRPF31* gene in patients with retinitis pigmentosa *(23)*. Both *RA_1* and *RA_2* are densely filled with various TEs (**Fig. S2** and **Table S7**), which are known to influence chromatin organization *(24)*, potentially influencing *TMEM106B* transcription.

An additional SV in a TR directly overlapping distal-enhancer, EH38E2534480 (**Fig. 2C** and **2E),** distinctly characterizes the *Hb* haplotype. This TR, referred to as $TR_{12157476}$, contains 13 unique motifs heavily composed of $(AAGG)_n$-like motifs (**Fig. 2E)**, previously linked with neurological syndromes, including Friedreich's Ataxia *(25)* and CANVAS syndrome *(26)*. We found 7 unique allele sequences for $TR_{12157476}$, mainly driven by variability in the total number of $(AAGG)_n$ motifs **(Fig. 2E)**. The shortest alleles (< 203 bp, AF = ~11%) were strongly enriched with the *Hb* haplotype (q-value = $4.29x10^{-137}$), while the remaining alleles (≥ 203 bp, AF = ~89%) occur in the other haplotypes.

Finally, an SV in a TR located in the 3' UTR of *TMEM106B*, referred to as $TR_{12237804}$, partially characterizes the Hb haplotype. $TR_{12237804}$ is mainly composed of a $(CA)_n$ motif, but contains a total of 16 unique allele sequences, with total lengths ranging from 36 to 80 bp (**Fig. 2F**). Although allele lengths were generally found across all haplotypes, the *Hb* haplotype was relatively enriched with the longest alleles (>67 bp) (q-value = $2.44x10^{-8}$), while the Ha haplotype was enriched with the shorter alleles (q-value = $1.09x10^{-2}$). Notably, TDP-43, an mRNA regulator, is

known to bind to mRNA molecules with expanded $(CA)_n$ motif sequences *(27, 28)*. Hence, different lengths of the $TR_{12237804}$ may provide differential sensitivity to TDP-43 interaction.

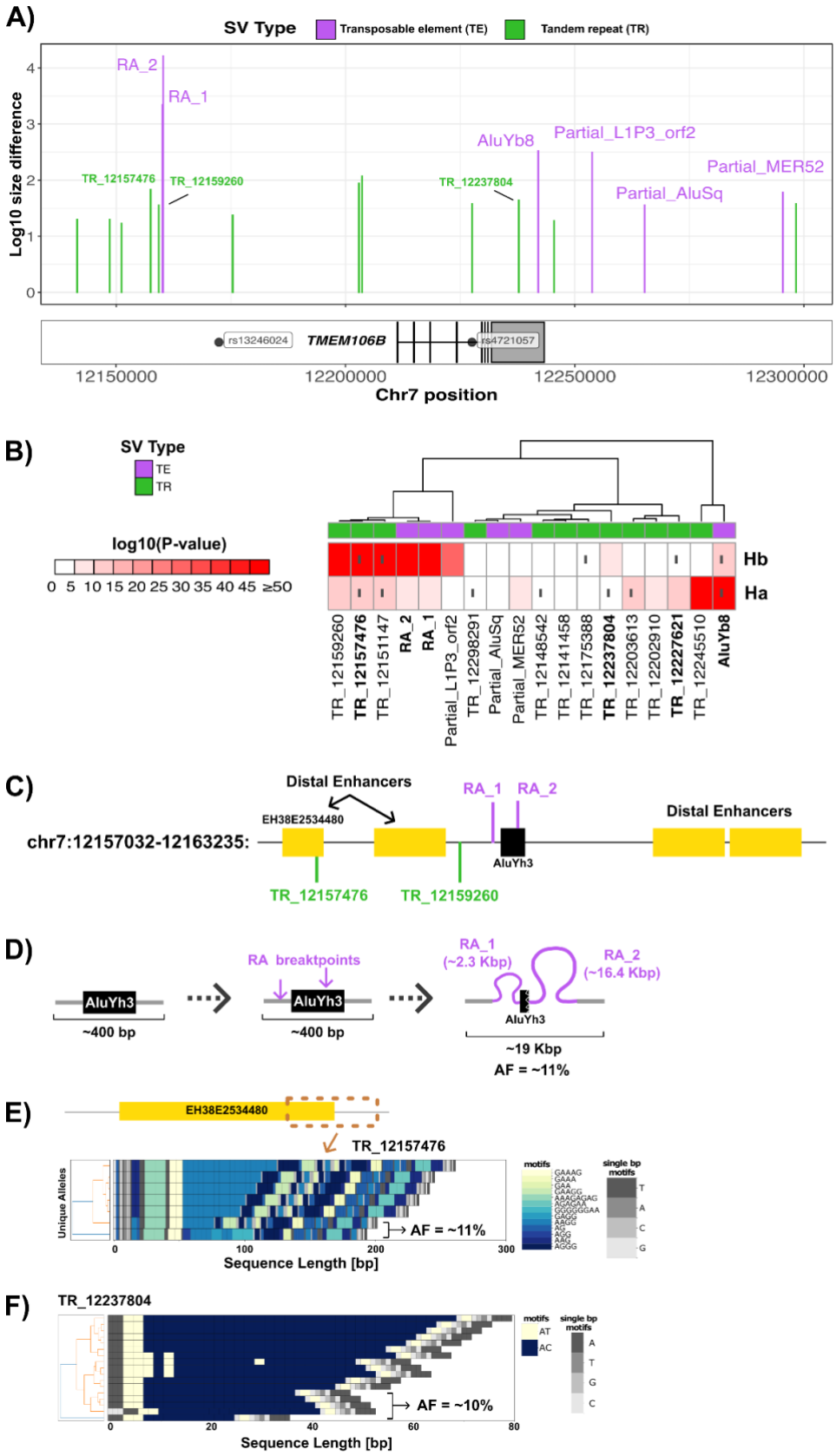

A)
SV Type
Transposable element (TE)
Tandem repeat (TR)
Log10 size difference
RA_2
RA_1
TR_12157476
TR_12159260
TR_12237804
AluYb8
Partial_L1P3_orf2
Partial_AluSq
Partial_MER52
rs13246024
TMEM106B
rs4721057
12150000
12200000
12250000
12300000
Chr7 position
B)
SV Type
TE
TR
log10(P-value)
0 5 10 15 20 25 30 35 40 45 ≥50
Hb
Ha
TR_12159260
TR_12157476
TR_12151147
RA_2
RA_1
Partial_L1P3_orf2
TR_12298291
Partial_AluSq
Partial_MER52
TR_12148542
TR_12141458
TR_12175388
TR_12237804
TR_12203613
TR_12202910
TR_12227621
TR_12245510
AluYb8
C)
Distal Enhancers
EH38E2534480
RA_1
RA_2
Distal Enhancers
chr7:12157032-12163235:
AluYh3
TR_12157476
TR_12159260
D)
RA breakpoints
RA_1
(~2.3 Kbp)
RA_2
(~16.4 Kbp)
AluYh3
~400 bp
~400 bp
AluYh3
~19 Kbp
AF = ~11%
E)
EH38E2534480
TR_12157476
Unique Alleles
motifs
single bp motifs
AF = ~11%
0
100
200
300
Sequence Length [bp]
F)
TR_12237804
motifs
AT
AC
single bp motifs
AF = ~10%
0
20
40
60
80
Sequence Length [bp]

**Figure 2. A total of 18 common structural variants (SVs) characterise the *TMEM106B* locus. (A-F)** Systematic identification of structural variants (SVs) within 100 Kbp upstream and downstream of the TMEM106B genic sequence revealed 18 SVs with allele frequencies ≥ 2.5% and >15 bp (see Table S4) from 493 ADC and CHC long-read sequenced genomes. **(A)** Top: Size differences (log10-scale) between the largest and smallest allele-sequences for each SV. SV positions are relative to GRCh38 chr7 coordinates. Bottom: TMEM106B canonical transcript (same as Fig. 1A), with black dots marking the locations of two AD-associated GWAS SNPs (see Fig. 1). **(B)** Association between SV allele-sequence lengths (columns) and Ha/Hb haplotypes (rows). White-to-red gradient represents strength of association (dark red, $p < 1x10^{-50}$; white, $p > 0.05$). Associations with a '-' indicate shorter SV alleles are linked to HA/HB, while the absence of '-' indicates the inverse (i.e. shorter SV alleles linked to Ha/Hb). Six SVs influencing *TMEM106B* transcription/translation are highlighted in bold (see Results). **(C)** 6.2 Kbp region (~49 Kbp upstream of the TMEM106B transcription start site) containing RA_1/RA_2 SVs (purple), two TRs (green), distal enhancers (yellow) as reported in the UCSC Genome Browser, and a human-specific AluYh3 element (black). **(D)** Diagram of RA_1/RA_2, revealing the differential presence of ~19 Kbp genomic sequence. **(E)** Visualization of $TR_{12157476}$, which partially spans distal-enhancer EH38E2534480. Seven unique allele-sequences identified across genomes (Y-axis) with length variation (X-axis) driven by an $(AAGG)_n$ motif (light blue). Shortest alleles (bottom rows, ~11% frequency) are in LD with RA_1/RA_2 (see Fig. 3). **(F)** Visualization of a 3' UTR repeat, $TR_{12237804}$. Seventeen unique allele-sequences with length variability driven by an $(AC)_n$ motif (dark blue). Shortest alleles (~10% frequency) are partially in LD with RA_1/RA_2.

## Associations of haplotype combinations support a protective role for SVs in the *Hb* haplotype, and a neutral role for the rs3173615-G (Serine) allele

Next, we performed long-read haplotype-phasing to validate the above haplotype-specific SV-enrichments. In addition to the 493 genomes from AD patients and CHCs, we matched the long-read sequences of 10 CHC blood-derived genomes with long-read sequencing of their brain-derived DNA genomes, and with the blood-derived genomes from one of their offspring. Together, our sample comprised 513 unique genomes, a total of 1,026 *TMEM106B* sequences.

We were able to completely phase 848 *TMEM106B* sequences, consisting of 396 AD sequences (198 genomes), and 452 centenarian sequences (211 genomes) **(Fig. 3A and Supplemental Methods**). Clustering analysis of the four major haplotype combinations showed that individuals in our cohorts inherit four distinct haplotype combinations: *HAB*, *HAb*, *HaB*, and *Hab* (**Fig. 3B).** Supporting this, the haplotype combinations are also observed in imputed phased-haplotypes from array genotype data of the 5,792 CHC, NDC, and ADC genomes, and all combinations are common (frequency ≥ 8.3%), with the exception of *HAb* (frequency ≤ 1%) **(Table S4)**.

Notably, long-read phasing and imputed phased-haplotypes showed that the *Hab* combination was nested, as ~90% of individuals in both datasets carrying *Hb* also carried *Ha*. This indicates that individuals that carry the *TMEM106B* alleles previously associated with protection (*Ha*) can be divided into two separate groups: *HaB* (33.7% frequency) and *Hab* (9.8% frequency) **(Fig. 3B, Table S4)**. Although both combinations carry the rs3173615-G (185-Serine) allele, only *Hab* carries the distal-enhancer SVs **(Fig. 3B)**. We hence explored how the addition of the SVs altered the protective effect of the 185-Serine allele. To do this, we investigated whether the 'risk' of becoming a CHC differed between carriers of the *HaB* (i.e. Serine only) and *Hab* (i.e. Serine and SVs) haplotype combinations, by performing a sensitivity analysis via separate joint logistic regression models relative to each haplotype combination (**Fig. 4 and Table S8, see Methods**)

When comparing CHCs with NDCs, we found a stepwise relationship between the effect-size contributions of each haplotype combination; such that *HAB* carried the most neurodegenerative risk, and *Hab* provided the most protective effect (**Fig. 4 and Table S8)**. Specifically, the *Hab* (i.e. SVs and 185-Serine) haplotype combination was associated with a 1.21-fold *higher* chance of becoming a CHC (OR = 1.21, CI = [0.915-1.59]) relative to *HaB* (i.e. 185-Serine only). Conversely, the *HaB* combination was associated with a 1.21-fold *decreased* chance of becoming a CHC (OR = 0.825, CI = [0.629-1.09]) relative to the *Hab* (i.e. SVs and 185-Serine) (**Fig. 4 and Table S8**). These results emphasize the substantial protective effect of the distal-enhancer SVs over the 185-Serine allele, as most of the neurodegenerative protection in CHCs is conferred by *Hab*. These findings remain consistent when comparing CHCs with ADCs, but with relatively lower effect-sizes (**Fig. 4 and Table S8**). Overall, these results indicate that the distal-enhancer SVs in *Hab* are the primary contributors to neurodegenerative protection from *TMEM106B* in CHCs, while the rs3173615-G (Serine) allele alone appears neutral or provides, at most, a modest additive protective effect.

In agreement to previous findings, the rs3173615-C (Threonine) allele and AluYb8 element are in full linkage, and distinctly characterized the risk-increasing *HA* haplotype and *HAB* combination, suggesting that the AluYb8 element may contribute to increased risk for neurodegeneration **(Fig. 3B)**. However, we also observed the less common *HAb* haplotype combination in 8 genomes (~1% frequency), which combines the distal-enhancer SVs (presence of *RA_1* and *RA_2*, and contractions of $TR_{12157476}$) along with the AluYb8 element **(Fig. 3B)**. Due to its low frequency, we could not determine to what extent this combination alters the risk of TMEM106B aggregation or the risk of neurodegenerative disease.

Finally, we note that these associations may differ in non-European populations due to differences in *TMEM106B* haplotype organization. For example, while rs1990620 and rs3173615 are in complete LD in genomes with European ancestry, LD is reduced in genomes with African ancestry *(29)*. Phasing variants in the same 109 Kbp *TMEM106B* region across 47 long-read sequenced genomes from diverse populations *(18)* revealed novel haplotype organizations (**Fig. S3** and **Supplemental Results**). Specifically, two unique haplotypes in African ancestry include combinations of rs3713615, the AluYb8 element, and the upstream SVs in the distal-enhancers, not generally seen in European genomes **(Fig. S3)**. These novel haplotypes may have distinct regulatory effects on *TMEM106B* neurodegenerative risk. Long-read sequencing across the global population will further uncover *TMEm106B* haplotype diversity.

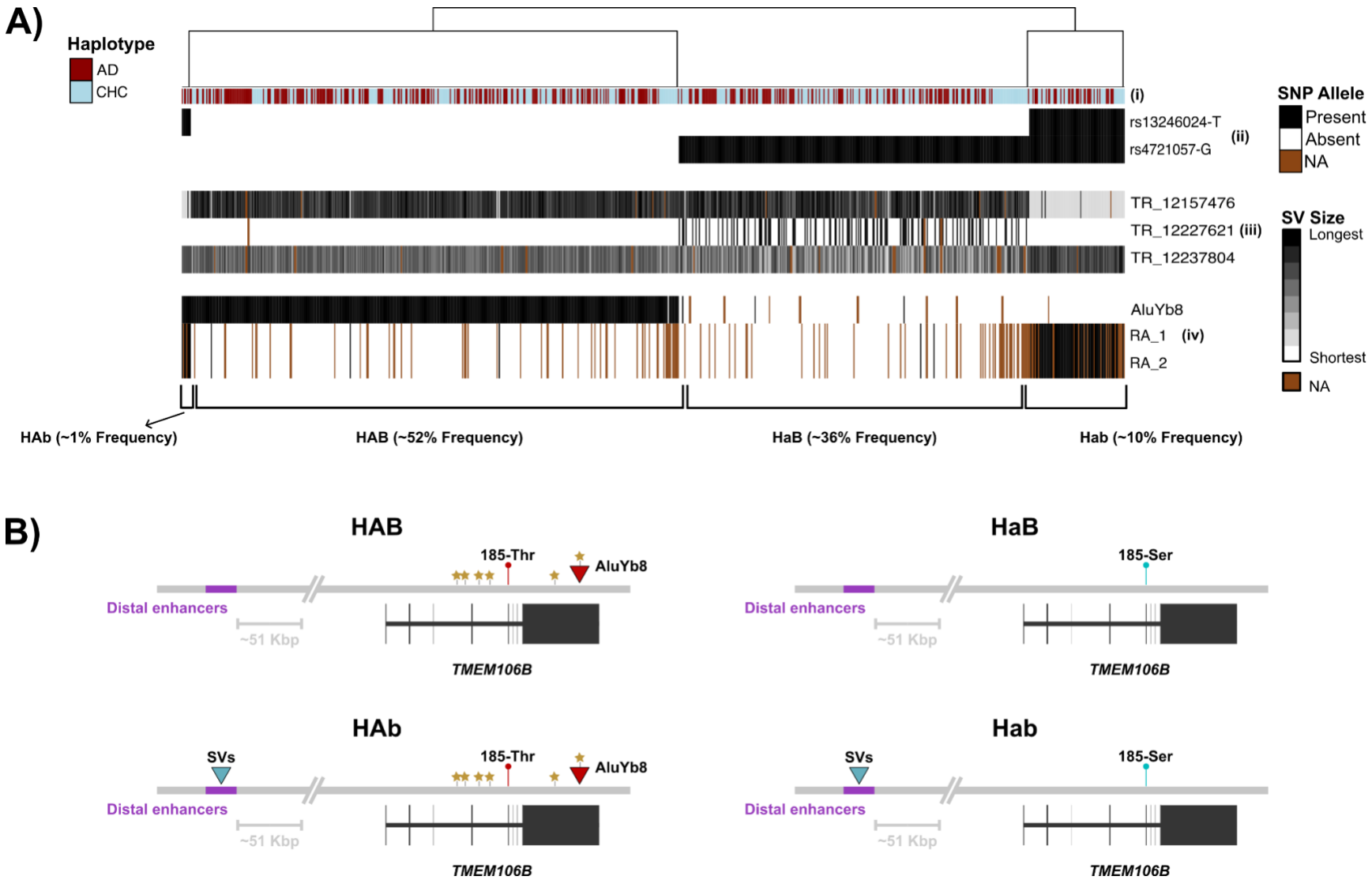



**Fig 3. Long-read phasing of all genetic variants in the *TMEM106B* locus reveals the existence of four major haplotype combinations**. **(A)** Long-read phasing of all common SNPs and SVs resulted in 852 complete haplotypes in the *TMEM106B* locus (see Results). Shown is a dendrogram of these haplotypes based on hierarchical-clustering of the rs4721057 and rs13246024 alleles

defining the *Ha* and *Hb* haplotypes, respectively; with the reciprocal alleles defining *HA* and *HB*, respectively. Each column represents a haplotype **(i)** originating from either a CHC (blue rectangles) or ADC (maroon rectangles). The upper three rows **(ii)** indicate presence (black) or absence (white) of the protective alleles of the haplotype marker variants from Fig. 1: rs4721057 and rs13246024. The middle three rows **(iii)** indicate the lengths of three tandem repeats (TR) using a colour-gradient from white to black, representing the shortest and longest allele within each SV, respectively. Finally, the bottom three rows **(iv)** are the same as (iii), but show the lengths of three TE-based SVs (see Fig. 2). For both SNPs and SVs, brown rectangles represent unknown genotypes due to low genotyping and/or phasing-quality. Hierarchical clustering of the haplotypes revealed four major clusters representing a haplotype combination, with their respective frequencies annotated at the bottom of the figure. **(B)** Illustration of the SVs characterizing the four major haplotype combinations along with the locations of the ~19 Kbp genomic rearrangement and contraction of $TR_{12157476}$ (referred to as SVs), the AluYb8 element, the 185-Threonine/Serine allele, and haplotype-specific CpG sites (gold stars) identified in Fig. 4 (see Results).

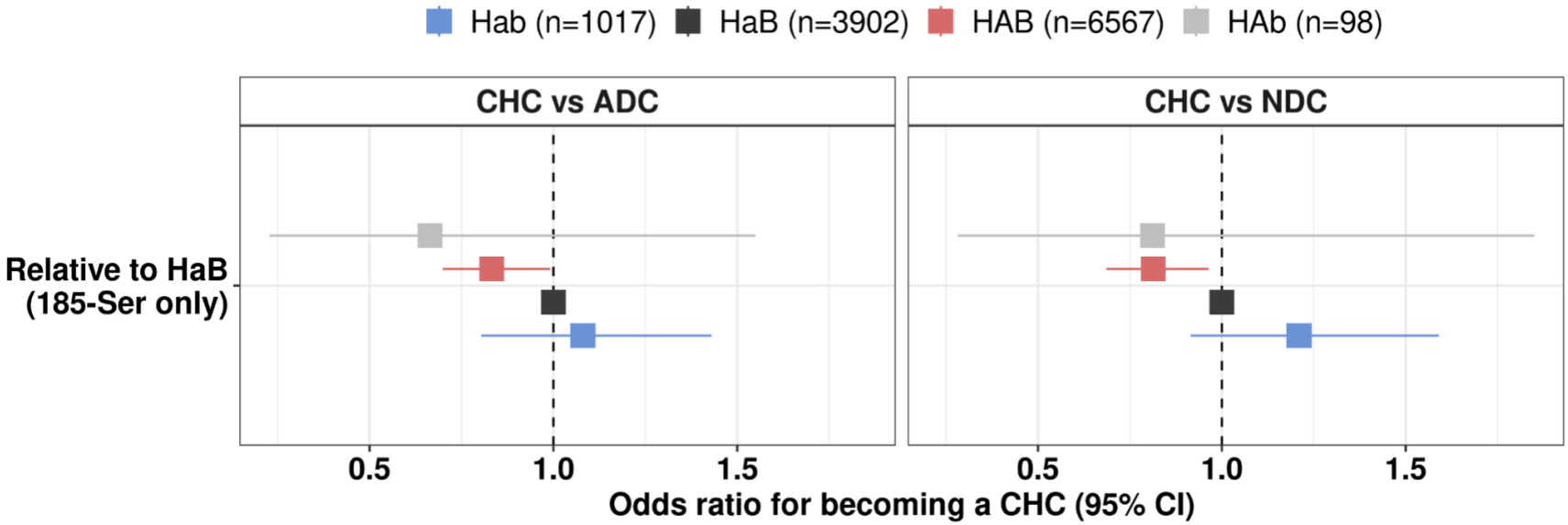



**Fig. 4. Joint logistic regression models reveal that the *Hab* haplotype combination (SVs + 185-Serine) confers the protective effect for becoming a CHC, while HaB (185-Serine only) is neutral.** The vast majority of *TMEM106B* neurodegenerative-associated risk-alleles are contained in the HAB haplotype combination, which carries the rs3173615-C (185-Threonine) allele and AluYb8 element. In contrast, the protective alleles of *TMEM106B* can be separated into two groups: carriers of HaB containing only the rs3173615-G (185-Serine) allele; and carriers of Hab containing both the185-Serine allele and distinct SVs within a cluster of distal-enhancers upstream of *TMEM106B* (Fig. 2 and 3). Given these distinctions, we investigated the relative protective effects conferred by the 185-Serine and SVs by calculating the 'risk' for becoming a CHC relative to HaB and Hab haplotype combinations. The figure shows a joint logistic regression model (Y-axis), with the respective odds ratio (X-axis) for each haplotype combination (coloured rectangles; with *n* corresponding to the total number of combinations observed across all three cohorts), along with the 95% confidence interval (CI, colored lines). The model shows the odds ratios relative to the *HaB* haplotype combination (i.e. 185-Serine only), comparing CHC with ADCs (left) and CHC with NDCs (right). Model coefficients are detailed in Table S8.

## The *Hab* haplotype combination are associated with lower *TMEM106B* C-terminal peptides

We investigated if the protective association of the *Hab* haplotype combination also led to lower abundance of soluble TMEM106B protein abundance in the brain. To do this, we performed mass spectrometry-based proteomic analysis of 182 temporal cortex tissues from Ganz *et al. (30)*, consisting of 55 CHC, 42 NDC, and 85 ADC samples (**Fig. 5**) with similar haplotype (combination) frequencies in the full cohorts and (**Table S9**). We focused on previously reported *EDAYDGVTSENMR* peptide that tags the N-terminal domain (residues 15-27); and the *SAYVSYDVQKR* peptide that tags the C-terminal part of TMEM106B (residues 130-140), representing the fibril core known to aggregate in human brains *(6–8)* (**Fig. 5A**). A joint linear regression analysis of all 182 tissues together revealed a similar stepwise relationship as observed in the genetic associations of the haplotype combinations (**Fig. 5C**): relative to the *HaB* haplotype combination (i.e. Serine only), the risk-increasing *HAB* combination was associated with 1.07-fold higher abundance of the C-terminal peptide (CI = [0.99-1.15]). While the *Hab* combination (i.e. SVs and Serine) was associated with a 0.91-fold lower abundance of the C-terminal peptide (CI = [0.81-1.02]). This is in line with the risk estimates of the haplotype combinations observed in GWAS and centenarians, supporting a protective role for the SVs in the *Hab* haplotype combination, and neutral role for the 185-Serine allele.

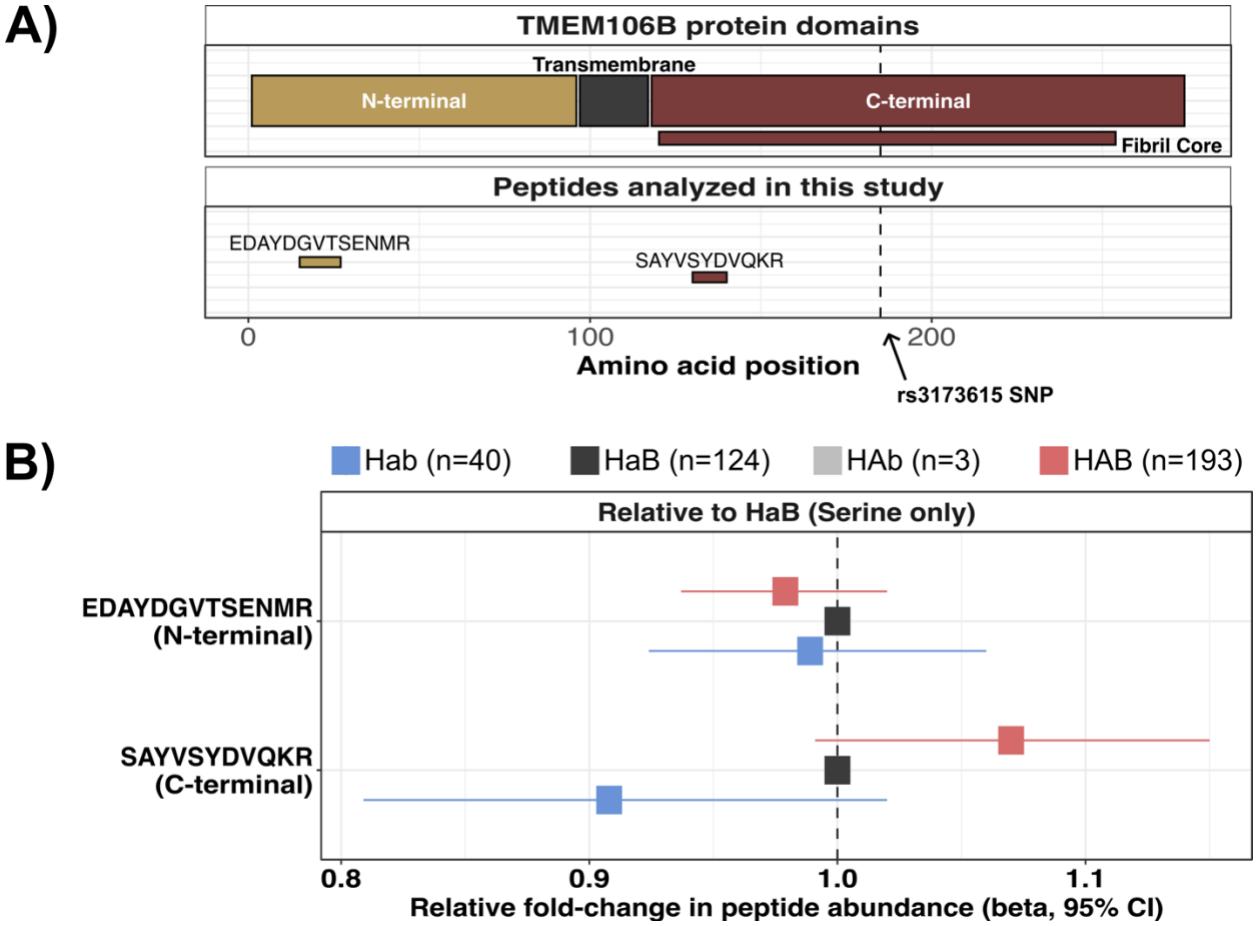



**Fig. 5. Proteomic analysis of 182 brain tissues reveals that the *Hab* haplotype combination (i.e. SVs + 185-Serine) leads to lower TMEM106B aggregation than HaB (i.e. 185-Serine only).** We used the results of mass spectrometry-based proteomic analysis of 182 temporal

cortex tissues representing a subset of CHC, NDC, and ADC samples (Table 1) from Ganz *et al.* 2023 *(30)*. **(A)** The top panel illustrates the TMEM106B protein domains along with the location of the fibril core, with the X-axis indicating the amino acid position. Similar to Lee et al. 2023 *(6)*, we focused on two peptides tagging the N- and C-terminal domains, illustrated in the bottom panel. **(B)** A joint linear regression model as similarly performed in Fig. 4, indicating the relative fold-change in peptide abundance associated with each haplotype combination *relative* to *HaB* (i.e. 185-Serine only). The legend indicates the total number of instances (*n*) a haplotype combination was observed across the 182 tissues. Note that *HAb* is included in the model but not in the figure as there were only 3 instances and hence has a large CI; model coefficients are detailed in Table S10.

**The risk-haplotype of *TMEM106B* have higher DNA-methylation in blood and brain**

With the long-read sequencing data, we were able to explore differential DNA methylation patterns at the CpG sites across the four haplotypes. TEs, such as the AluYb8 element in *HA* haplotype (**Fig. 3**), reportedly increase DNA methylation in surrounding regions due to evolutionary pressure to suppress their transcriptional activation *(31)*, preventing further propagation of the TE to maintain overall genome stability *(32)*. This prompted us to explore differential methylation signals of the CpG sites in the four *TMEM106B* haplotypes.

We identified 675 unique CpG sites throughout the *TMEM106B* locus (**Fig. 6**, see **Methods**). Of these, 24 were specific to the AluYb8 sequence, and hence present in all AuYb8 carriers (**Supplemental Results**). We observed high variability in the methylation frequency of the AluYb8-bound CpGs sites, including those in RNA polymerase III recognition sequences, suggesting that the AluYb8 element may be transcriptionally active in some individuals (**Supplemental Results**).

We found that 54/651 non-AluYb8-bound CpG sites were significantly differentially methylated across AluYb8 zygosity, leading to overall higher DNA methylation in the *TMEM106B* gene for carriers of the HA haplotype (**Fig. 6C**, **Table S11,** see **Methods**). Based on LD and haplotype phasing analysis, we found that 40/54 Cpg sites varied across individuals **(Fig. 6D)**: 16 CpG sites were specific to the HA haplotype, enriched in the *TMEM106B* introns 3 and 4; while 5 CpGs were specific to the HB haplotype, located nearby the AluYh3 element replacing distal-enhancer SVs upstream of *TMEM106B* (**Fig. 2C**); and 19 were variably found across multiple haplotypes (**Table S12**). Hence, age-related DNA methylation changes in these CpG sites may lead to alternative processing of *TMEM106B* mRNA molecules, such as alternative splicing *(33)*. Indeed, quantitative trait loci (eQTLs) of the common CpG marker variants in GTEx shows that the HA haplotype is associated with retention of intron 3 (strongest association in Esophagus - Mucosa tissue, p-value = $1.11 \times 10^{-12}$).

The remaining 14/54 Cpg sites were common (AF ≥ 2.5%) across all individuals regardless of haplotype ('Universal', **Fig. 6D**). Of these, we found that the CpG sites closer to the AluYb8 element had relatively higher methylation for AluYb8 carriers, suggesting active pressure to suppress AluYb8 transcription *(32).* However, we also found higher methylation for non-AluYb8 carriers upstream of *TMEM106B* near the location of the distal-enhancer SVs, suggesting an active regulatory role of the distal-enhancer SVs.

Finally, we did not observe significant differences in CpG methylation between CHC and ADCs (**Fig. S4**). However, while methylation levels generally overlap between blood and brain and across zygosities of the AluYb8 element (**Fig. S5)**, we observed several tissue-specific DNA

methylation patterns independent of AluYb8 zygosity (**Fig. S6)**. This includes both, brain-specific decreased methylation in introns 1, 3 and 4; and brain-specific increased methylation in introns 3, 4, and the 3' UTR (**Fig. S6**). These tissue-specific methylation patterns may further drive alternative splicing of *TMEM106B*.

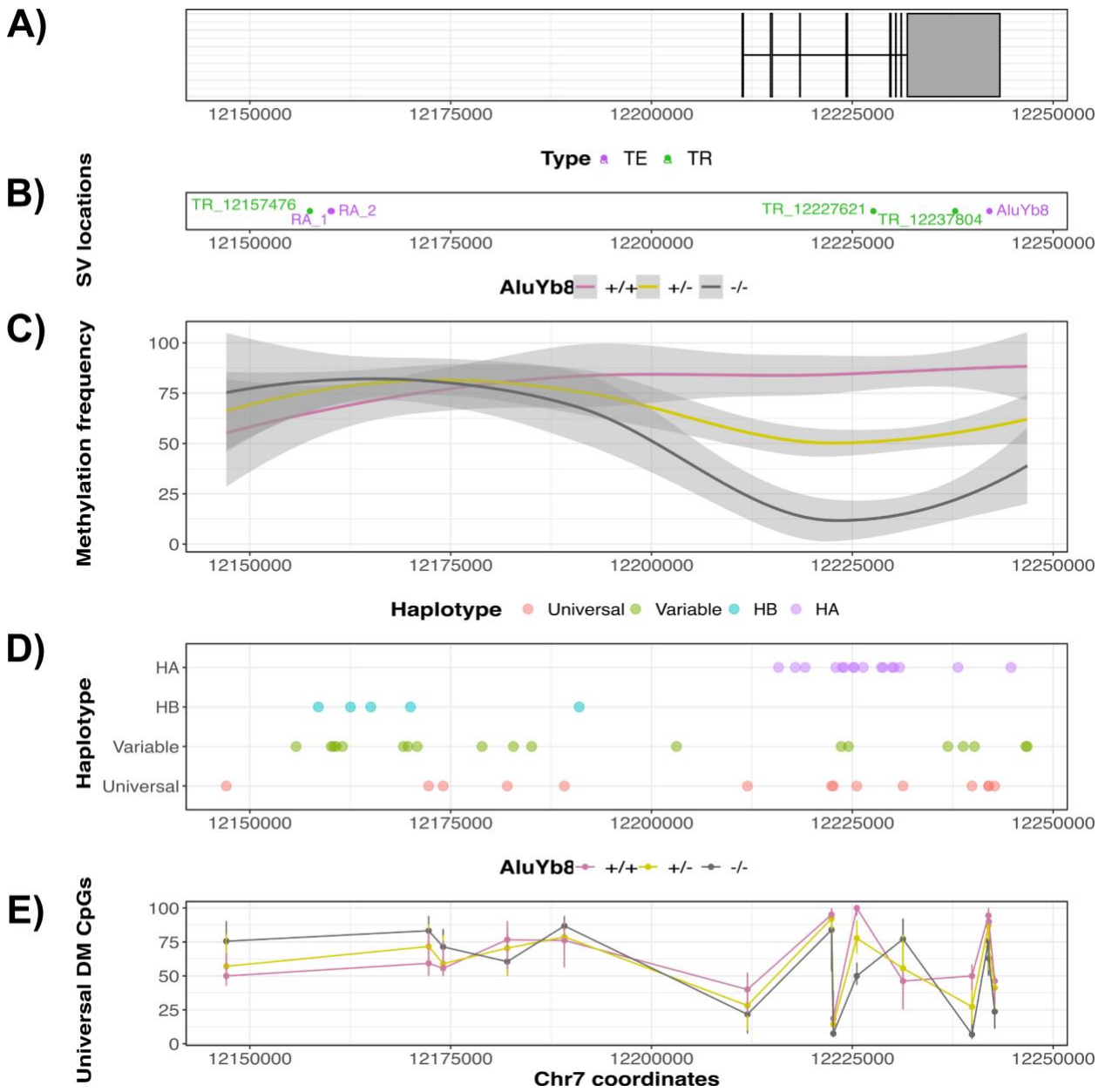



**Fig. 6. The risk haplotype of the *TMEM106B* locus has higher DNA methylation levels across 513 ADC and CHC genomes in both peripheral blood and brain tissue. (A)** *TMEM106B* canonical transcript as in Fig. 1. **(B)** Location of six SVs from Fig. 2 and Fig. 3 relative to GRCh38 coordinates (X-axis). **(C)** DNA methylation frequency (Y-axes) across the *TMEM106B* locus (X-axis) from peripheral blood derived from long-read sequencing methylation signals, partitioned by homozygous carriers of the AluYb8 element (pink), heterozygous carriers (orange), and homozygous non-carriers (grey). This was based on local polynomial regression using the

54 significantly differentially methylated (DM) CpG sites in Fig. 6D, with standard errors depicted by the grey shadow on each line. **(D)** The location of 54 DM CpG sites, with colors representation different annotations: purple indicates CpG sites specific to HA haplotype; blue specific to HB; red indicates variable CpG sites that are variably methylated across individuals (i.e. part of common variants) but not specific to haplotypes in Fig. 1; and red indicates CpG sites that are universally present across all individuals (AF ≥ 2.5%). **(E)** Methylation of DM CpG sites that are universally present, where each point is the median methylation frequency and the lines indicate the interquartile range (25-75%), such that the colors correspond to AluYb8 zygosity as in Fig. 6A.

## DISCUSSION

We demonstrate that two independent linkage disequilibrium (LD) blocks in *TMEM106B* associate with Alzheimer's disease (AD), generating four haplotypes: *HA*, *Ha*, *HB*, and *Hb*. The *HA* haplotype carries risk alleles, including rs3173615-C (185-Threonine) and an AluYb8 element, which we show increases DNA methylation and TMEM106B protein aggregation. In contrast, the *Ha* haplotype carries protective alleles, including rs3173615-G (185-Serine). While *HB* lacks notable disease-related genetic elements, *Hb* includes various structural variants (SVs), such as a 19 Kbp rearrangement within upstream cluster of distal-enhancers. In our cohorts, these haplotypes form four different combinations: *HAB*, *HAb*, *HaB*, *Hab*. Notably, both *HaB* and *Hab* carry the protective 185-Serine allele, but *Hab* carry various distinct SVs in upstream distal-enhancers. We find that cognitively healthy centenarians are relatively more enriched with the protective *Hab* combination than *HaB*, and mass spectrometry-based proteomic analysis shows that *Hab* is also linked to lower C-terminal peptide abundance in temporal cortex brain tissue relative to *HaB*. These results suggest that the 185-Serine allele is neutral, while the *Hb*-specific SVs confer neurodegenerative protective effects in *TMEM106B*, potentially through changes in chromatin organization that influence *TMEM106B* gene regulation.

Additional haplotypes exist in *TMEM106B,* but it is unclear to what extent they influence genetic susceptibility to neurodegeneration. For example, long-read sequencing indicated that the Ha haplotype includes an intronic bi-allelic TR in some individuals. Moreover, the total length of the $(CA)_n$ TR in the 3' UTR that can vary widely across haplotypes, which is notable given that TDP-43 has binding affinity to expanded $(CA)_n$ repeats *(27, 28)*. Therefore, *TMEM106B* mRNA molecules with expanded $(CA)_n$ repeats may be sensitive to processing by TDP-43, potentially influencing *TMEM106B* translation. Hence, in addition to a possible mechanistic link between TDP-43 and the AluYb8 element with the risk haplotype *(34)*, there could also be a mechanistic link between TDP-43 and expanded $(CA)_n$ TR alleles.

Together, these results suggest that a decrease in TDP-43 activity may interfere with *TMEM106B* regulation. Indeed, the levels of TMEM106B and TDP-43 protein aggregates are significantly correlated in human brains *(8)*. Overall, a larger GWAS study of an AD endophenotype-driven GWAS *(3)*, preferably with the addition of representative SNPs for each of the SVs/lengths, and additional non-European haplotypes, may elucidate whether the TRs are part of additional haplotypes that are independently associated with neurodegenerative risk. Of particular interest are rs74835425-A allele (tagging the bi-allelic intronic TR expansion), and the rs73301035-A allele (tagging the $(CA)_n$ TR contraction) (see **Supplemental Results**).

Given the unique genetic features of the four haplotypes architectures, our study also provides preliminary indications that the *TMEM106B* is regulated via multiple cell-specific, and age-related factors. Functional experiments are needed to validate the mechanistic link between the haplotypes and regulation of the *TMEM106B* gene, but the hypothesized mechanistic effect on disease risk, associated with each haplotype is motivated by previous functional characterizations of genomic elements similar to the SVs that define them. These include: **(i)** chromatin reorganization via the SVs in the distal-enhancer cluster upstream of *TMEM106B*, potentially leading to differential activation of these enhancers *(24, 25, 35)*; **(ii)** translational regulation of *TMEM106B* via interactions of non-coding RNAs with the AluYb8 element (Staufen-mediated decay) *(36)*; **(iii)** AluYb8 transcriptional reactivation leading to *TMEM106B* transcriptional interference due to competing RNA polymerases *(37)*; and **(iv)** *TMEM106B* alternative splicing due to tissue-specific and age-related demethylation of CpG sites exclusive to the *TMEM106B* risk haplotypes. Each regulatory mechanism likely differs across tissues, may be dependent on the activity levels of TDP-43, and is likely further exacerbated by other neuropathological factors.

Finally, we show that unique *TMEM106B* haplotypes exist throughout the human population, including haplotypes with a unique combination of coding and regulatory variants in individuals with African ancestry. Therefore, the above modes of mechanisms may differ in other non-European populations. We note that the marker variants of the haplotypes and combinations are not universally applicable across other ancestries, highlighting the advantages of direct genotyping SVs with long-read sequencing. Nevertheless, larger long-read whole-genome datasets are needed to reveal the full extent of haplotype diversity in the *TMEM106B* locus.

In conclusion, the regulatory and functional consequences of the four haplotypes identified above elucidate a complex genetic etiology of *TMEM106B*. This includes age-related and tissue-specific effects that may influence aggregation of TMEM106B protein, as well as overall integrity of the endolysosomal system *(9)*. The fine-mapping of the *TMEM106B* locus also underscores that the incorporation of large SVs and methylation patterns is a prerequisite to capture the full spectrum of heritable risk factors associated with the pathophysiology underlying neurodegenerative diseases. This provides broader implications for understanding the etiology underlying the genetic susceptibility conferred by other disease-associated genetic loci.

## METHODS

### Sample populations

An overview of the samples used in this study is shown in (**Table 1**). Briefly, 346 cognitively healthy centenarians (referred to as CHCs) in this study, along with the 10 centenarian brain samples and 10 peripheral blood samples of centenarian offspring, are individuals from the 100-plus Study *(20)*. The 2,281 Alzheimer disease (AD) patients in this study are from the Amsterdam Dementia Cohort *(21)* as well as the Netherlands Brain Bank *(38)* (referred to as ADCs). The 3,165 ADC-aged matched non-demented individuals (referred to as NDCs) were previously described *(4)*. 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.

### Selection and association of GWAS variants and haplotypes in the *TMEM106B* locus

We analyzed GWAS variants within 100 Kbp upstream and downstream of *TMEM106B* coding sequence (chr7:12111294-12343367), using an updated AD-GWAS summary statistics from Bellenguez *et al*. 2023 *(19)*, involving ADGC, Bonn, CHARGE, EADB, EADI, FinnGen, GERAD, GR@ACE/DEGESCO, and PGC-ALZ consortia (**Supplemental Materials)**. Pairwise linkage disequilibrium (LD) values ($r^2$) were calculated with PLINK (v2.00a6LM) using the CHC, ADC, and NCD cohorts (see section above), and variants were clumped using a conservative $r^2$ threshold of 0.15. We retained the sentinel SNPs of clumps with genome-wide significance in the AD-GWAS (p-value < $5x10^{-8}$). This revealed two LD-blocks, which we refer to as *L1* and *L2*. To avoid overestimation in downstream analysis, we performed hierarchical clustering of all SNPs in strong LD ($r^2$ >= 0.95) with the sentinel SNP from each LD-block, and selected the medoid SNP of each block as the marker SNP. Joint effects analysis with *GCTA-COJO* (v1.94.1) confirmed their independent associations. Logistic regression models were then performed to evaluate the association of the dosages of the marker SNPs with CHC 'risk'. Corrected for the first five principal components, using phased imputed genotypes for CHC, ADC, and NDC cohorts. Both marginal models (individual marker SNP dosages) and joint models (dosages of both SNPs) were applied, comparing CHC versus NDC and CHC versus ADC. False discovery rate (FDR) correction was applied to all p-values using the Benjamin-Hochberg procedure (q = 5%). Statistical analysis was conducted using the *glm* package in R.

### PacBio HiFi long-read whole-genome sequencing and data processing

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

After sequencing, raw reads were collected and analyzed through an in-house pipeline (freely available at https://github.com/holstegelab/snakemake_pipeline; **Supplemental Methods**). In addition, we used publicly available HiFi-based 47 whole-genomes from the Human Pangenome Research Consortium, which were similarly processed through our in-house pipeline (**Supplemental Methods**).

**Variant calling and LD analysis**

Single nucleotide polymorphisms (SNPs) were identified in both the ADC and CHC long-read sequencing data using *deepvariant* (v1.1.0), and joint-called using *GLnexus* (v1.4.1-0-g68e25e5) using '--config Deepvariant' parameter. The same procedure was performed separately for the HPRC genomes and the five CHC genomes with representative *TMEM106B* haplotypes.

Structural variants (SVs) across AD, centenarians, and HPRC genomes were identified and genotyped via an in-house local assembly pipeline based on *sniffles2 (39)* (version 2.0.6) and *otter* (v1.0) *(17)* (**Supplemental Methods**). This generated locally assembled allele sequence(s) per candidate SV per genome. For the genomic rearrangement (*RA_1* and *RA_2* SVs), we first performed *de novo* assembly of all genomes with *hifiasm* (v0.16.1-r375), and aligned the resulting phased contigs to the GRCh38 genome. The genomic rearrangements in aligned phased-contigs that fully span the *TMEM106B* target region (chr7:12146000-12255000) were then genotyped with *otter* using the whole-genome alignment mode, 'wgat', generating allele sequence(s) per genomic rearrangement per SV.

We developed a joint-genotyping procedure to generically compare and associate allele sequences of SVs across cohorts **(Supplemental Results)**. This enables us to perform rapid joint-genotyping of SVs, scalable across >1 million SV-regions and large collections of long-read sequenced genomes (**Supplemental Results**). We used this procedure to joint-genotype all allele-sequences of each SV in the *TMEM016B* locus, generating a SV-based gVCF across the ADC and CHC genomes; and separately for the HPRC and representative CHC genomes.

We approximated SV-to-SNP LD by performing linear regression to associate the allele-lengths of all SVs with the dosage of SNPs. Briefly, we defined dosage of an SV in a genome as the sum of total lengths based on its genotype (see above). We then used a generalised linear model to regress SNP dosage using total SV-length as the variable, corrected for the first five principal components: *Dosage_SNP ~ Dosage_SV + PC1 + PC2 + PC3 + PC4 + PC5*. All p-values were FDR-adjusted for multiple-testing correction (q-value = 0.05).

Finally, the corresponding allele-sequence of all unique genotypes of all SVs were annotated analysed with DFAM (release 3.7) *(40)* to annotate transposable elements (TEs).

**Phasing, clustering, and association of *TMEM106B* haplotype (combinations)**

We performed long-read phasing of SNPs and SVs in the *TMEM106B* target region (chr7:12146000-12255000) with hiphase (v1.4.2) *(41)* by joining *deepvariant* and *otter* variant calls (**Supplemental Methods**). This enabled us to represent CHC, ADC, and HPRC genomes as a haplotype-block matrix where columns represented SNP and SV genotypes and rows representing haplotype blocks per genome. *TMEM106B* haplotypes were then clustered to investigate possible risk and protective sub-haplotypes, using the genotypes of the rs472105 and rs13246024 GWAS SNPs. The resulting matrix was clustered using R's *hclust* package (v4.4), and visualized using *pheatmap* (v1.0.12).

We performed four different analyses to investigate differential association of the TMEM106B haplotype combinations. First we used imputed phased data of the CHC, ADC, and NDC genomes (see section above) to determine the zygosity (i.e. dosage) of each haplotype combination for each individual based on the allele definitions in Table 2. To assess enrichment in CHCs, we use separate marginal logistic regression models based on the zygosities of each haplotype combination separately, corrected for the first five principal components using the glm package in R:

(1) $log(\frac{P(Y=1)}{P(Y=0)}) = \beta_0 + \beta_1 D_h + \beta_2 PC_1 + \beta_3 PC_2 + \beta_4 PC_3 + \beta_5 PC_4 + \beta_6 PC_5$

In association (1), $P(Y = 1)$ indicates if a genome is a CHC; and D indicates the dosage (i.e. 1, 0, or 2) of haplotype combination h, such that $h \in H$ where $H = \{HAB, HAb, HaB, Hab\}$, while $PC_1$-$PC_5$ are the first five principal components.

We then performed a joint logistic regression model to investigate if the HaB and Hab haplotype combinations are differentially enriched in CHCs:

(2) $log(\frac{P(Y=1)}{P(Y=0)}) = \beta_0 + \beta_1 D_{HAb} + \beta_2 D_{HaB} + \beta_3 D_{Hab} +$
$\beta_4 PC_1 + \beta_5 PC_2 + \beta_6 PC_3 + \beta_7 PC_4 + \beta_8 PC_5$

In association (2), the dosage of each of the respective haplotype combinations is an independent variable. The exception is the HAB haplotype combination, which is excluded (i.e. reference combination), and hence, all effect-sizes in (2) are relative to not carrying the HAB haplotype.

Finally, we performed two separate joint logistic regression models to investigate the relative protective effect-size contribution from the HaB (i.e. 185-Serine only) and Hab (i.e. 185-Serine and SVs) haplotype combinations. To do this, we performed a similar joint logistic regression association as above, but included HAB (i.e. 185-Threonine and AluYb8) and varied the *TMEM106B* protective alleles:

(3) $log(\frac{P(Y=1)}{P(Y=0)}) = \beta_0 + \beta_1 D_{HAB} + \beta_2 D_{HAb} + \beta_3 D_{HaB} +$
$\beta_4 PC_1 + \beta_5 PC_2 + \beta_6 PC_3 + \beta_7 PC_4 + \beta_8 PC_5$

(4) $log(\frac{P(Y=1)}{P(Y=0)}) = \beta_0 + \beta_1 D_{HAB} + \beta_2 D_{HAb} + \beta_3 D_{Hab} +$
$\beta_4 PC_1 + \beta_5 PC_2 + \beta_6 PC_3 + \beta_7 PC_4 + \beta_8 PC_5$

Association (3) excludes dosage of the Hab combination, and hence the protective *TMEM106B* alleles are only represented by HaB (i.e. 185-Serine only). Similarly, association (4) excludes HaB, and hence the protective alleles are represented by Hab (i.e. 185-Serine and SVs).

All four association analyses were performed between CHCs and NDCs, as well as CHCs and ADCs. For each association, we FDR-corrected all p-values from the CHC, NDC, and ADC comparisons together.

**Mass spectrometry-based proteomic analysis**

Mass spectrometry-based proteomic analysis of 182 temporal cortex tissues (Table 1) was performed as previously reported *(30)* using the Bruker TIMS-TOF-pro-HT mass spectrometer (**Supplemental Methods**). We focused on the normalized abundance of two peptides, *EDAYDGVTSENMR* and *SAYVSYDVQKR*, which tag the N- and C-terminal domains of TMEM106B, respectively; as reported by Lee *et al.* 2023 *(6)* and *(8)* Marks *et al.* 2024. Finally, we performed a joint linear regression association across all 182 tissues to investigate how the haplotype combinations contributed to TMEM106B brain aggregation:

(5) $Y_p = \beta_0 + \beta_1 D_{HAB} + \beta_2 D_{HAb} + \beta_3 D_{Hab} + \beta_4 PC_1 + \beta_5 PC_2 + \beta_6 PC_3 + \beta_7 PC_4 + \beta_8 PC_5 + \epsilon$

In association (5), $Y_p$ is the normalised abundance of peptide *p*, where *p* is either *SAYVSYDVQKR* (C-terminal) or *EDAYDGVTSENMR* (N-terminal). Note that the dosage of *HaB* (i.e. 185-Serine only) is excluded in the model, and hence, all fold-changes associated with each haplotype combination are *relative* to *HaB*.

**Profiling methylation in TMEM106B locus**

Methylation frequency of CpG sites were calculated using PacBio's *pg-CpG-tools* (version 2.3.1, https://github.com/PacificBiosciences/pb-CpG-tools) with default parameters, along with the "--modsites-mode reference" parameter on aligned BAM files to GRCh38. Rare CpG sites that were only present in less than 2% of genomes (n = 10) were removed. The methylation values were associated with binarized dosage of the AluYb8 genotype (see previous section) using the *glm* package in R. P-values were FDR-adjusted for multiple-testing correction (q-value = 0.05). For determining significant differentially methylated CpG sites throughout the *TMEM106B* locus based on AluYb8 element zygosity, we retained CpG's with q-value ≤ 0.05, and median methylation difference ≥ 10% between homozygous AluYb8 carriers and homozygous non-carriers. Sequence variation in CpG sites were analyzed by identifying corresponding dinucleotide sequences based on both LD analysis with haplotype marker SNPs, and phased sequences of *TMEM106B* (see section above). RNA polymerase III and transcription binding recognition sequences within the AluYb8 element were cross-referenced against published annotations from Conti *et al.* 2015 *(42).*

**List of Supplementary Materials**

Consortia members and affiliations can be found in Supplemental Materials
Supplemental Results and Methods can be found in Supplemental Materials
Fig. S1-S6 can be found in Supplemental Materials
Table S1-S9 can be found in Supplemental Tables

**REFERENCES**

1. C. Bellenguez, F. Küçükali, I. E. Jansen, L. Kleineidam, S. Moreno-Grau, N. Amin, A. C. Naj, R. Campos-Martin, B. Grenier-Boley, V. Andrade, P. A. Holmans, A. Boland, V. Damotte, S. J. van der Lee, M. R. Costa, T. Kuulasmaa, Q. Yang, I. de Rojas, J. C. Bis, A. Yaqub, I. Prokic, J. Chapuis, S. Ahmad, V. Giedraitis, D. Aarsland, P. Garcia-Gonzalez, C. Abdelnour, E. Alarcón-Martín, D. Alcolea, M. Alegret, I. Alvarez, V. Álvarez, N. J. Armstrong, A. Tsolaki, C. Antúnez, I. Appollonio, M. Arcaro, S. Archetti, A. A. Pastor, B. Arosio, L. Athanasiu, H. Bailly, N. Banaj, M. Baquero, S. Barral, A. Beiser, A. B. Pastor, J. E. Below, P. Benchek, L. Benussi, C. Berr, C. Besse, V. Bessi, G. Binetti, A. Bizarro, R. Blesa, M. Boada, E. Boerwinkle, B. Borroni, S. Boschi, P. Bossù, G. Bråthen, J. Bressler, C. Bresner, H. Brodaty, K. J. Brookes, L. I. Brusco, D. Buiza-Rueda, K. Bûrger, V. Burholt, W. S. Bush, M. Calero, L. B. Cantwell, G. Chene, J. Chung, M. L. Cuccaro, Á. Carracedo, R. Cecchetti, L. Cervera-Carles, C. Charbonnier, H.-H. Chen, C. Chillotti, S. Ciccone, J. A. H. R. Claassen, C. Clark, E. Conti, A. Corma-Gómez, E. Costantini, C. Custodero, D. Daian, M. C. Dalmasso, A. Daniele, E. Dardiotis, J.-F. Dartigues, P. P. de Deyn, K. de Paiva Lopes, L. D. de Witte, S. Debette, J. Deckert, T. del Ser, N. Denning, A. DeStefano, M. Dichgans, J. Diehl-Schmid, M. Diez-Fairen, P. D. Rossi, S. Djurovic, E. Duron, E. Düzel, C. Dufouil, G. Eiriksdottir, S. Engelborghs, V. Escott-Price, A. Espinosa, M. Ewers, K. M. Faber, T. Fabrizio, S. F. Nielsen, D. W. Fardo, L. Farotti, C. Fenoglio, M. Fernández-Fuertes, R. Ferrari, C. B. Ferreira, E. Ferri, B. Fin, P. Fischer, T. Fladby, K. Fließbach, B. Fongang, M. Fornage, J. Fortea, T. M. Foroud, S. Fostinelli, N. C. Fox, E. Franco-Macías, M. J. Bullido, A. Frank-García, L. Froelich, B. Fulton-Howard, D. Galimberti, J. M. García-Alberca, P. García-González, S. Garcia-Madrona, G. Garcia-Ribas, R. Ghidoni, I. Giegling, G. Giorgio, A. M. Goate, O. Goldhardt, D. Gomez-Fonseca, A. González-Pérez, C. Graff, G. Grande, E. Green, T. Grimmer, E. Grünblatt, M. Grunin, V. Gudnason, T. Guetta-Baranes, A. Haapasalo, G. Hadjigeorgiou, J. L. Haines, K. L. Hamilton-Nelson, H. Hampel, O. Hanon, J. Hardy, A. M. Hartmann, L. Hausner, J. Harwood, S. Heilmann-Heimbach, S. Helisalmi, M. T. Heneka, I. Hernández, M. J. Herrmann, P. Hoffmann, C. Holmes, H. Holstege, R. H. Vilas, M. Hulsman, J.

Humphrey, G. J. Biessels, X. Jian, C. Johansson, G. R. Jun, Y. Kastumata, J. Kauwe, P. G. Kehoe, L. Kilander, A. K. Ståhlbom, M. Kivipelto, A. Koivisto, J. Kornhuber, M. H. Kosmidis, W. A. Kukull, P. P. Kuksa, B. W. Kunkle, A. B. Kuzma, C. Lage, E. J. Laukka, L. Launer, A. Lauria, C.-Y. Lee, J. Lehtisalo, O. Lerch, A. Lleó, W. Longstreth, O. Lopez, A. L. de Munain, S. Love, M. Löwemark, L. Luckcuck, K. L. Lunetta, Y. Ma, J. Macías, C. A. MacLeod, W. Maier, F. Mangialasche, M. Spallazzi, M. Marquié, R. Marshall, E. R. Martin, A. M. Montes, C. M. Rodríguez, C. Masullo, R. Mayeux, S. Mead, P. Mecocci, M. Medina, A. Meggy, S. Mehrabian, S. Mendoza, M. Menéndez-González, P. Mir, S. Moebus, M. Mol, L. Molina-Porcel, L. Montrreal, L. Morelli, F. Moreno, K. Morgan, T. Mosley, M. M. Nöthen, C. Muchnik, S. Mukherjee, B. Nacmias, T. Ngandu, G. Nicolas, B. G. Nordestgaard, R. Olaso, A. Orellana, M. Orsini, G. Ortega, A. Padovani, C. Paolo, G. Papenberg, L. Parnetti, F. Pasquier, P. Pastor, G. Peloso, A. Pérez-Cordón, J. Pérez-Tur, P. Pericard, O. Peters, Y. A. L. Pijnenburg, J. A. Pineda, G. Piñol-Ripoll, C. Pisanu, T. Polak, J. Popp, D. Posthuma, J. Priller, R. Puerta, O. Quenez, I. Quintela, J. Q. Thomassen, A. Rábano, I. Rainero, F. Rajabli, I. Ramakers, L. M. Real, M. J. T. Reinders, C. Reitz, D. Reyes-Dumeyer, P. Ridge, S. Riedel-Heller, P. Riederer, N. Roberto, E. Rodriguez-Rodriguez, A. Rongve, I. R. Allende, M. Rosende-Roca, J. L. Royo, E. Rubino, D. Rujescu, M. E. Sáez, P. Sakka, I. Saltvedt, Á. Sanabria, M. B. Sánchez-Arjona, New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics* **54**, 412–436 (2022).

2. V. M. Van Deerlin, P. M. A. Sleiman, M. Martinez-Lage, A. Chen-Plotkin, L.-S. Wang, N. R. Graff-Radford, D. W. Dickson, R. Rademakers, B. F. Boeve, M. Grossman, S. E. Arnold, D. M. A. Mann, S. M. Pickering-Brown, H. Seelaar, P. Heutink, J. C. Van Swieten, J. R. Murrell, B. Ghetti, S. Spina, J. Grafman, J. Hodges, M. G. Spillantini, S. Gilman, A. P. Lieberman, J. A. Kaye, R. L. Woltjer, E. H. Bigio, M. Mesulam, S. al-Sarraj, C. Troakes, R. N. Rosenberg, C. L. White, I. Ferrer, A. Lladó, M. Neumann, H. A. Kretzschmar, C. M. Hulette, K. A. Welsh-Bohmer, B. L. Miller, A. Alzualde, A. L. De Munain, A. C. McKee, M. Gearing, A. I. Levey, J. J. Lah, J. Hardy, J. D. Rohrer, T. Lashley, I. R. A. Mackenzie, H. H. Feldman, R. L. Hamilton, S. T. Dekosky, J. Van Der Zee, S. Kumar-Singh, C. Van Broeckhoven, R. Mayeux, J. P. G. Vonsattel, J. C. Troncoso, J. J. Kril, J. B. J. Kwok, G. M. Halliday, T. D. Bird, P. G. Ince, P. J. Shaw, N. J. Cairns, J. C. Morris, C. A. McLean, C. DeCarli, W. G. Ellis, S. H. Freeman, M. P. Frosch, J. H. Growdon, D. P. Perl, M. Sano, D. A. Bennett, J. A. Schneider, T. G. Beach, E. M. Reiman, B. K. Woodruff, J. Cummings, H. V. Vinters, C. A. Miller, H. C. Chui, I. Alafuzoff, P. Hartikainen, D. Seilhean, D. Galasko, E. Masliah, C. W. Cotman, M. T. Tuñón, M. C. C. Martínez, D. G. Munoz, S. L. Carroll, D. Marson, P. F. Riederer, N. Bogdanovic, G. D. Schellenberg, H. Hakonarson, J. Q. Trojanowski, V. M.-Y. Lee, Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. *Nat Genet* **42**, 234–239 (2010).

3. L. M. P. Shade, Y. Katsumata, E. L. Abner, K. Z. Aung, S. A. Claas, Q. Qiao, B. A. Heberle, J. A. Brandon, M. L. Page, T. J. Hohman, S. Mukherjee, R. P. Mayeux, L. A. Farrer, G. D. Schellenberg, J. L. Haines, W. A. Kukull, K. Nho, A. J. Saykin, D. A. Bennett, J. A. Schneider, The National Alzheimer's Coordinating Center, W. A. Kukull, A. J. Saykin, D. A. Bennett, The Alzheimer's Disease Genetics Consortium, E. L. Abner, S. Mukherjee, R. P. Mayeux, L. A. Farrer, G. D. Schellenberg, J. L. Haines, D. A. Bennett, J. A. Schneider, M. T. W. Ebbert, P. T. Nelson, D. W. Fardo, GWAS of multiple neuropathology endophenotypes identifies new risk loci and provides insights into the genetic risk of dementia. *Nat Genet* (2024), doi:10.1038/s41588-024-01939-9.

4. N. Tesi, S. Van Der Lee, M. Hulsman, N. M. Van Schoor, M. Huisman, Y. Pijnenburg, W. M. Van Der Flier, M. Reinders, H. Holstege, Cognitively healthy centenarians are genetically protected against Alzheimer's disease. *Alzheimer's & Dementia* **20**, 3864–3875 (2024).

5. M. Zhang, A. B. Ganz, S. Rohde, L. Lorenz, A. J. M. Rozemuller, K. Van Vliet, M. Graat, S. A. M. Sikkes, M. J. T. Reinders, P. Scheltens, M. Hulsman, J. J. M. Hoozemans, H. Holstege, The correlation between neuropathology levels and cognitive performance in centenarians. *Alzheimer's & Dementia* , alz.13087 (2023).

6. J. Y. Lee, D. J. Harney, J. D. Teo, J. B. Kwok, G. T. Sutherland, M. Larance, A. S. Don, The major TMEM106B dementia risk allele affects TMEM106B protein levels, fibril formation, and myelin lipid homeostasis in the ageing human hippocampus. *Mol Neurodegeneration* **18**, 63 (2023).

7. C. T Vicente, J. Perneel, S. Wynants, B. Heeman, M. Van Den Broeck, M. Baker, S. Cheung, J. Faura, I. R. A. Mackenzie, R. Rademakers, C-terminal TMEM106B fragments in human brain correlate with disease-associated *TMEM106B* haplotypes. *Brain* , awad133 (2023).

8. J. D. Marks, V. E. Ayuso, Y. Carlomagno, M. Yue, T. W. Todd, Y. Hao, Z. Li, Z. T. McEachin, A. Shantaraman, D. M. Duong, L. M. Daughrity, K. Jansen-West, W. Shao, A. Calliari, J. G. Bejarano, M. DeTure, B. Rawlinson, M. C. Casey, M. T. Lilley, M. H. Donahue, V. M. Jawahar, B. F. Boeve, R. C. Petersen, D. S. Knopman, B. Oskarsson, N. R. Graff-Radford, Z. K. Wszolek, D. W. Dickson, K. A. Josephs, Y. A. Qi, N. T. Seyfried, M. E. Ward, Y.-J. Zhang, M. Prudencio, L. Petrucelli, C. N. Cook, TMEM106B core deposition associates with TDP-43 pathology and is increased in risk SNP carriers for frontotemporal dementia. *Sci. Transl. Med.* **16**, eadf9735 (2024).

9. H.-S. Jiao, P. Yuan, J.-T. Yu, TMEM106B aggregation in neurodegenerative diseases: linking genetics to function. *Mol Neurodegeneration* **18**, 54 (2023).

10. A.-S. Cabron, U. Borgmeyer, J. Richter, H. Peisker, K. Gutbrod, P. Dörmann, A. Capell, M. Damme, Lack of a protective effect of the Tmem106b "protective SNP" in the Grn knockout mouse model for frontotemporal lobar degeneration. *acta neuropathol commun* **11**, 21 (2023).

11. A. M. Nicholson, N. A. Finch, A. Wojtas, M. C. Baker, R. B. Perkerson, M. Castanedes-Casey, L. Rousseau, L. Benussi, G. Binetti, R. Ghidoni, G.-Y. R. Hsiung, I. R. Mackenzie, E. Finger, B. F. Boeve, N. Ertekin-Taner, N. R. Graff-Radford, D. W. Dickson, R. Rademakers, TMEM106B p.T185S regulates TMEM106B protein levels: implications for frontotemporal dementia. *J. Neurochem.* **126**, 781–791 (2013).

12. A. Chemparathy, Y. L. Guen, Y. Zeng, J. Gorzynski, T. Jensen, N. Kasireddy, L. Talozzi, M. E. Belloy, I. Stewart, A. D. Gitler, A. D. Wagner, E. Mormino, V. W. Henderson, T. Wyss-Coray, E. Ashley, M. D. Greicius, *A 3' UTR Deletion Is a Leading Candidate Causal Variant at the* TMEM106B *Locus Reducing Risk for FTLD-TDP* (Neurology, 2023; http://medrxiv.org/lookup/doi/10.1101/2023.07.06.23292312).

13. A. Rodney, K. Karanjeet, K. Benzow, M. D. Koob, A common Alu insertion in the 3'UTR of *TMEM106B* is associated with risk of dementia. *Alzheimer's & Dementia* **20**, 5071–5077 (2024).

14. A. Salazar, N. Tesi, L. Knoop, Y. Pijnenburg, S. Van Der Lee, S. Wijesekera, J. Krizova, M. Hiltunen, M. Damme, L. Petrucelli, M. Reinders, M. Hulsman, H. Holstege, An AluYb8 retrotransposon characterises a risk haplotype of TMEM106B associated in neurodegeneration (2023), doi:10.1101/2023.07.16.23292721.

15. M. D. Gallagher, M. Posavi, P. Huang, T. L. Unger, Y. Berlyand, A. L. Gruenewald, A. Chesi, E. Manduchi, A. D. Wells, S. F. A. Grant, G. A. Blobel, C. D. Brown, A. S. Chen-Plotkin, A Dementia-Associated Risk Variant near TMEM106B Alters Chromatin Architecture and Gene Expression. *The American Journal of Human Genetics* **101**, 643–663 (2017).

16. C. Depienne, J.-L. Mandel, 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *The American Journal of Human Genetics* **108**, 764–785 (2021).

17. N. Tesi, A. Salazar, Y. Zhang, S. Van Der Lee, M. Hulsman, L. Knoop, S. Wijesekera, J. Krizova, A.-F. Schneider, M. Pennings, K. Sleegers, E.-J. Kamsteeg, M. Reinders, H. Holstege, Characterising tandem repeat complexities across long-read sequencing platforms with TREAT and otter (2024), doi:10.1101/2024.03.15.585288.

18. W.-W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, H. J. Abel, S. Buonaiuto, X. H. Chang, H. Cheng, J. Chu, V. Colonna, J. M. Eizenga, X. Feng, C. Fischer, R. S. Fulton, S. Garg, C. Groza, A. Guarracino, W. T. Harvey, S. Heumos, K. Howe, M. Jain, T.-Y. Lu, C. Markello, F. J. Martin, M. W. Mitchell, K. M. Munson, M. N. Mwaniki, A. M. Novak, H. E. Olsen, T. Pesout, D. Porubsky, P. Prins, J. A. Sibbesen, J. Sirén, C. Tomlinson, F. Villani, M. R. Vollger, L. L. Antonacci-Fulton, G. Baid, C. A. Baker, A. Belyaeva, K. Billis, A. Carroll, P.-C. Chang, S. Cody, D. E. Cook, R. M. Cook-Deegan, O. E. Cornejo, M. Diekhans, P. Ebert, S. Fairley, O. Fedrigo, A. L. Felsenfeld, G. Formenti, A. Frankish, Y. Gao, N. A. Garrison, C. G. Giron, R. E. Green, L. Haggerty, K. Hoekzema, T. Hourlier, H. P. Ji, E. E. Kenny, B. A. Koenig, A. Kolesnikov, J. O. Korbel, J. Kordosky, S. Koren, H. Lee, A. P. Lewis, H. Magalhães, S. Marco-Sola, P. Marijon, A. McCartney, J. McDaniel, J. Mountcastle, M. Nattestad, S. Nurk, N. D. Olson, A. B. Popejoy, D. Puiu, M. Rautiainen, A. A. Regier, A. Rhie, S. Sacco, A. D. Sanders, V. A. Schneider, B. I. Schultz, K. Shafin, M. W. Smith, H. J. Sofia, A. N. Abou Tayoun, F. Thibaud-Nissen, F. F. Tricomi, J. Wagner, B. Walenz, J. M. D. Wood, A. V. Zimin, G. Bourque, M. J. P. Chaisson, P. Flicek, A. M. Phillippy, J. M. Zook, E. E. Eichler, D. Haussler, T. Wang, E. D. Jarvis, K. H. Miga, E. Garrison, T. Marschall, I. M. Hall, H. Li, B. Paten, A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

19. C. Bellenguez, Consensus meta-analysis of genome-wide association studies for Alzheimer's disease across the EADB, PGC-ALZ and IGAP consortia. *Alzheimer's & Dementia* **19**, e076317 (2023).

20. H. Holstege, N. Beker, T. Dijkstra, K. Pieterse, E. Wemmenhove, K. Schouten, L. Thiessens, D. Horsten, S. Rechtuijt, S. Sikkes, F. W. A. Van Poppel, H. Meijers-Heijboer, M. Hulsman, P. Scheltens, The 100-plus Study of cognitively healthy centenarians: rationale, design and cohort description. *Eur J Epidemiol* **33**, 1229–1249 (2018).

21. W. M. Van Der Flier, P. Scheltens, G. Perry, J. Avila, M. Tabaton, X. Zhu, Eds. Amsterdam Dementia Cohort: Performing Research to Optimize Care. *JAD* **62**, 1091–1111 (2018).

22. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics* **88**, 76–82 (2011).

23. Q. Lu, J. Yang, Y. Xiong, X. Zhu, M. Zhang, D. Zhou, A novel genomic rearrangement on chr19q13.42 leads to *PRPF31* -associated retinitis pigmentosa. *Clinical Exper Ophthalmology* **52**, 595–598 (2024).

24. H. A. Lawson, Y. Liang, T. Wang, Transposable elements in mammalian chromatin organization. *Nat Rev Genet* **24**, 712–723 (2023).

25. A. M. Gacy, G. M. Goellner, C. Spiro, X. Chen, G. Gupta, E. M. Bradbury, R. B. Dyer, M. J. Mikesell, J. Z. Yao, A. J. Johnson, A. Richter, S. B. Melançon, C. T. McMurray, GAA Instability in Friedreich's Ataxia Shares a Common, DNA-Directed and Intraallelic Mechanism with Other Trinucleotide Diseases. *Molecular Cell* **1**, 583–593 (1998).

26. A. Cortese, R. Simone, R. Sullivan, J. Vandrovcova, H. Tariq, W. Y. Yau, J. Humphrey, Z. Jaunmuktane, P. Sivakumar, J. Polke, M. Ilyas, E. Tribollet, P. J. Tomaselli, G. Devigili, I. Callegari, M. Versino, V. Salpietro, S. Efthymiou, D. Kaski, N. W. Wood, N. S. Andrade, E. Buglo, A. Rebelo, A. M. Rossor, A. Bronstein, P. Fratta, W. J. Marques, S. Züchner, M. M. Reilly, H. Houlden, Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat Genet* **51**, 649–658 (2019).

27. P.-H. Kuo, L. G. Doudeva, Y.-T. Wang, C.-K. J. Shen, H. S. Yuan, Structural insights into TDP-43 in nucleic-acid binding and domain interactions. *Nucleic Acids Research* **37**, 1799–1808 (2009).

28. Z. R. Grese, A. C. Bastos, L. D. Mamede, R. L. French, T. M. Miller, Y. M. Ayala, Specific RNA interactions promote TDP-43 multivalent phase separation and maintain liquid properties. *EMBO Reports* **22**, e53632 (2021).

29. Y. Katsumata, D. W. Fardo, L. M. P. Shade, Alzheimer's Disease Genetics Consortium, J. D. Bowen, P. K. Crane, G. P. Jarvik, C. D. Keene, E. B. Larson, W. C. McCormick, S. M. McCurry, S. Mukherjee, N. W. Kowall, A. C. McKee, R. A. Honig, S. Lawrence, J. P. Vonsattel, J. Williamson, S. Small, J. R. Burke, C. M. Hulette, K. A. Welsh-Bohmer, M. Gearing, J. J. Lah, A. I. Levey, T. S. Wingo, L. G. Apostolova, M. R. Farlow, B. Ghetti, A. J. Saykin, S. Spina, M. S. Albert, C. G. Lyketsos, J. C. Troncoso, M. P. Frosch, R. C. Green, J. H. Growdon, B. T. Hyman, R. E. Tanzi, H. Potter, D. W. Dickson, N. Ertekin-Taner, N. R. Graff-Radford, J. E. Parisi, R. C. Petersen, R. Duara, J. D. Buxbaum, A. M. Goate, M. Sano, A. V. Masurkar, T. Wisniewski, E. H. Bigio, M. Mesulam, S. Weintraub, R. Vassar, J. A. Kaye, J. F. Quinn, R. L. Woltjer, L. L. Barnes, D. A. Bennett, J. A. Schneider, L. Yu, V. Henderson, K. B. Fallon, L. E. Harrell, D. C. Marson, E. D. Roberson, C. DeCarli, L.-W. Jin, J. M. Olichney, R. Kim, F. M. LaFerla, E. Monuki, E. Head, D. Sultzer, D. H. Geschwind, H. V. Vinters, M.-F. Chesselet, D. R. Galasko, J. B. Brewer, A. Boxer, A. Karydas, J. H. Kramer, B. L. Miller, H. J. Rosen, W. W. Seeley, J. M. Burns, R. H. Swerdlow, E. Abner, D. W. Fardo, L. J. Van Eldik, R. L. Albin, A. P. Lieberman, H. L. Paulson, S. E. Arnold, J. Q. Trojanowski, V. M. Van Deerlin, R. L. Hamilton, M. I. Kamboh, O. L. Lopez, J. T. Becker, C. Cao, A. Raj, A. G. Smith, H. C. Chui, C. A. Miller, J. M. Ringman, L. S. Schneider, T. D. Bird, J. A. Sonnen, C.-E. Yu, T. Grabowski, E. Peskind, M. Raskind, G. Li, D. W. Tsuang, S. Asthana, C. S. Atwood, C. M. Carlsson, M. A. Sager, N. A. Chin, S. Craft, N. J. Cairns, J. C. Morris, C. Cruchaga, S. Strittmatter, E. M. Reiman, T. G. Beach, M. J. Huentelman, J. Hardy, A. J. Myers, J. S. K. Kauwe, H. Hakonarson, D. Blacker, T. J. Montine, C. T. Baldwin, L. A. Farrer, G. Jun, K. L. Lunetta, W. S. Bush, J. L. Haines, A. J. Lerner, X. Zhou, S. Barral, C. Reitz, B. N. Vardarajan, R. Mayeux, G. W. Beecham, R. M. Carney, M. L. Cuccaro, J. R. Gilbert, K. L. Hamilton-Nelson, B. W. Kunkle, E. R. Martin, M. A. Pericak-Vance, J. M. Vance, L. B. Cantwell, A. P. Kuzma, J. Malamon, A. C. Naj, L. Qu, G. D. Schellenberg, O. Valladares, L.-S. Wang, Y. Zhao, J. B. Leverenz, P. L. De Jager, D. A. Evans, M. J. Katz, R. B. Lipton, B. F. Boeve, M. Allen, M. M. Carrasquillo, S. G. Younkin, W. A. Kukull, K. M. Faber, T. M. Foroud, V. Pavlik, P. Massman, E. Darby, M. Rodriguear, A. Khaleeq, D. R. Royall, A. Stevens, M. Ory, J. C. DeToledo, H. Wilms, K. Johnson, V. Perez, M. Hernandez, K. C. Wilhelmsen, J. Tilson, S.

Chasse, R. C. Barber, T. J. Fairchild, S. E. O'Bryant, J. Knebl, J. R. Hall, L. Johnson, D. Mains, L. Alvarez, A. Gamboa, D. Paydarfar, J. Bertelson, M. Woon, G. Ayres, A. Aguirre, R. Palmer, M. Polk, P. M. Adams, R. M. Huebinger, J. S. Reisch, R. N. Rosenberg, M. Cullum, B. Williams, M. Quiceno, L. Hynan, J. Smith, B. Davis, T. Nguyen, E. Rogaeva, P. S. George-Hyslop, P. T. Nelson, LATE-NC risk alleles (in *TMEM106B, GRN* , and *ABCC9* genes) among persons with African ancestry. *Journal of Neuropathology & Experimental Neurology* **82**, 760–768 (2023).

30. A. B. Ganz, M. Zhang, F. Koopmans, K. W. Li, S. S. M. Miedema, A. J. M. Rozemuller, M. Hulsman, Netherlands Brain Bank, P. Scheltens, J. J. M. Hoozemans, M. J. T. Reinders, A. B. Smit, H. Holstege, Proteomic profiling of aging brains identifies key proteins by which cognitively healthy centenarians defy their age by decades (2023), doi:10.1101/2023.11.30.23299224.

31. W. Zhou, G. Liang, P. L. Molloy, P. A. Jones, DNA methylation enables transposable element-driven genome expansion. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 19359–19366 (2020).

32. M. V. Almeida, G. Vernaz, A. L. K. Putman, E. A. Miska, Taming transposable elements in vertebrates: from epigenetic silencing to domestication. *Trends in Genetics* **38**, 529–553 (2022).

33. D. Kim, M. Shivakumar, S. Han, M. S. Sinclair, Y.-J. Lee, Y. Zheng, O. I. Olopade, D. Kim, Y. Lee, Population-dependent Intron Retention and DNA Methylation in Breast Cancer. *Molecular Cancer Research* **16**, 461–469 (2018).

34. O. H. Tam, N. V. Rozhkov, R. Shaw, D. Kim, I. Hubbard, S. Fennessey, N. Propp, D. Fagegaltier, B. T. Harris, L. W. Ostrow, H. Phatnani, J. Ravits, J. Dubnau, M. Gale Hammell, H. Phatnani, J. Kwan, D. Sareen, J. R. Broach, Z. Simmons, X. Arcila-Londono, E. B. Lee, V. M. Van Deerlin, N. A. Shneider, E. Fraenkel, L. W. Ostrow, F. Baas, N. Zaitlen, J. D. Berry, A. Malaspina, P. Fratta, G. A. Cox, L. M. Thompson, S. Finkbeiner, E. Dardiotis, T. M. Miller, S. Chandran, S. Pal, E. Hornstein, D. J. MacGowan, T. Heiman-Patterson, M. G. Hammell, Nikolaos. A. Patsopoulos, O. Butovsky, J. Dubnau, A. Nath, R. Bowser, M. Harms, E. Aronica, M. Poss, J. Phillips-Cremins, J. Crary, N. Atassi, D. J. Lange, D. J. Adams, L. Stefanis, M. Gotkine, R. Baloh, S. Babu, T. Raj, S. Paganoni, O. Shalem, C. Smith, B. Zhang, B. T. Harris, Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Reports* **29**, 1164-1177.e5 (2019).

35. T. Punga, M. Bühler, Long intronic GAA repeats causing Friedreich ataxia impede transcription elongation. *EMBO Mol Med* **2**, 120–129 (2010).

36. A. S. Chen-Plotkin, T. L. Unger, M. D. Gallagher, E. Bill, L. K. Kwong, L. Volpicelli-Daley, J. I. Busch, S. Akle, M. Grossman, V. Van Deerlin, J. Q. Trojanowski, V. M.-Y. Lee, *TMEM106B* , the Risk Gene for Frontotemporal Dementia, Is Regulated by the microRNA-132/212 Cluster and Affects Progranulin Pathways. *J. Neurosci.* **32**, 11213–11227 (2012).

37. L.-L. Chen, L. Yang, ALU ternative Regulation for Gene Expression. *Trends in Cell Biology* **27**, 480–490 (2017).

38. M. C. Rademaker, G. M. De Lange, S. J. M. C. Palmen, in *Handbook of Clinical Neurology*, (Elsevier, 2018), vol. 150, pp. 3–16.

39. M. Smolka, L. F. Paulin, C. M. Grochowski, M. Mahmoud, S. Behera, M. Gandhi, K. Hong, D. Pehlivan, S. W. Scholz, C. M. B. Carvalho, C. Proukakis, F. J. Sedlazeck, *Comprehensive Structural Variant Detection: From Mosaic to Population-Level* (Bioinformatics, 2022;

http://biorxiv.org/lookup/doi/10.1101/2022.04.04.487055).

40. J. Storer, R. Hubley, J. Rosen, T. J. Wheeler, A. F. Smit, The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* **12**, 2 (2021).

41. J. M. Holt, C. T. Saunders, W. J. Rowell, Z. Kronenberg, A. M. Wenger, M. Eberle, L. Cowen, Ed. HiPhase: Jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. *Bioinformatics* , btae042 (2024).

42. A. Conti, D. Carnevali, V. Bollati, S. Fustinoni, M. Pellegrini, G. Dieci, Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Research* **43**, 817–835 (2015).

**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.

**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).

**Author contributions**
Conceived the study: HH; Wrote the manuscript: AS, HH; Patient selection: NT, HH, Patient blood collection and sequencing, MG, MH, DD, SW, JK LK; Data management: AS, NT, MH, SvdL, Bioinformatic analysis: AS.

**Competing interests**
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.

**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/).