An AluYb8 mobile element characterises a risk haplotype of TMEM106B associated in neurodegeneration

Alex Salazar[1]*, Niccolò Tesi[1,2], Marc Hulsman[1,2], Lydian Knoop[1], Sven van der Lee[1,2,4], Sanduni Wijesekera[1], Jana Krizova[1], Marcel Reinders[2], and Henne Holstege[1,2,3,4,]**

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

Corresponding author:

Henne Holstege: h.holstege@amsterdamumc.nl

Abstract

Genome-wide association studies have identified a role for *TMEM106B* in frontotemporal lobar degeneration with TAR DNA binding protein inclusions (FTD-LD) and Alzheimer's disease (AD). However, the functional effect of the *TMEM106B* risk haplotype remains elusive. Based on long-read whole-genome sequencing data of 256 individuals, we identified an AluYb8 mobile element that is inserted in the 3'UTR of TMEM106B of the risk haplotype. When activated by demethylation, the AluY8b-element may affect *TMEM106B* transcription. Investigation of methylation levels in the risk haplotype suggests (partial) mobile element activation in older individuals. Furthermore, we found that methylation levels in 3' UTR of *TARDBP* in both AD and centenarian genomes may be reflective of a reduced Alu-element suppressing activity of the TDP-43 protein. Haplotype-phasing revealed four major haplotype groups in Europeans and African genomes, which may explain the reduced association of *TMEM106B* on disease risk in individuals from African ancestry. Finaly, we propose a mechanism for how TMEM106B, PGRN, TDP-43, and age interact in neurodegeneration via a negative-feedback loop that leads to a progressive decline of the endolysosomal system.

Introduction

TMEM106B likely plays a central role in neurodegenerative diseases. The gene encodes a 274 amino-acid transmembrane protein highly expressed in the central nervous system, especially in neurons and oligodendrocytes[1,2]. Although its functions are not entirely clear, TMEM106B localizes in the endolysosomal system, where it is involved in maintaining lysosome integrity [1,2]. Genome-wide association studies (GWAS) have highlighted the presence of two common haplotypes in *TMEM106B* (allele-frequency ~50%), one associated with increased risk of neurodegenerative diseases, and hence the other with a protective effect[2–10] (see Figure 1 A). The haplotypes are characterised by several common variants that are in strong-to-complete linkage disequilibrium (LD) with each other (see Figure 1 A and B). The rs1990622 variant was associated with protection against frontotemporal lobar degeneration with TAR DNA binding protein inclusions (FTDLD-TDP) (OR=0.61; $p = 1.08 \times 10^{-11}$)[3]. The rs13237518 variant was associated with protection against Alzheimer's disease (OR=0.96, $p = 4.90 \times 10^{-11}$)[11]. Notably, the association of the protective haplotype in *TMEM106B* with FTDLD-TDP is more pronounced in carriers of pathogenic *GRN* variants (OR=0.49, $p = 1.34 \times 10^{-9}$)[2,3], a gene encoding progranulin (PGRN)[12]. Since both genes are essential for lysosome function[12], especially in the degradation of misfolded proteins[13], their effects may converge in the impairment of lysosomal mechanisms.

Both the rs1990622 and the rs13237518 variants are in strong linkage with rs3173615, a non-synonymous variant (T185S) which substitutes a threonine residue (risk) for a serine residue (protective) in exon 6 of *TMEM106B*[2,9]. This substitution has become the focus of many studies aiming to explore the functional consequences of the protective haplotype of *TMEM106B*[2,5,14–18]. The T185S substitution is predicted to be 'benign' by Polyphen, SIFT and

REVEL (score of 0.0267), suggesting that the subsitution is unlikely to have major effect on TMEM106B function. The T185S substitution juxtaposes an *N*-glycosylation site at position 183, which has been proposed to possibly affect post-translational protein maturation[19]. However, *N*-glycosylation sites commonly occur at a N-X-S/T motif, suggesting that substituting the N_{183}-I_{184}-T_{185} motif to N_{183}-I_{184}-S_{185} motif is unlikely to have any effect[19]. Notably, there were no observable functional changes in a CRISPR-based knockin-mouse model homozygous for the T185S substitution[14].

Nevertheless, several studies hypothesized that T185S may alter *N*-glycosylation such that the risk haplotype has a higher aggregation potential[18–24]. Aggregate amyloid fibrils composed of TMEM106B protein were observed in the post-mortem brain of humans with advanced age, or brains from individuals suffering from neurodegenerative diseases[20–24]. They were specifically composed of the C-terminal fragment (CTF) of TMEM106B[20–24], and carriers of the risk haplotype had higher amounts of CTF compared to carriers of the protective haplotype[18]. Cryogenic-electron microscopy showed that the aggregated fibrils have three distinct structural polymorphisms: fold I is predominant in individuals homozygous for the risk allele threonine (TT) and heterozygous carriers (TS); fold II represents only a small fraction of TMEM106B CTF fibrils in carriers of the risk haplotype; and fold III is predominant in individuals homozygous for the protective allele serine (SS).[20–24]. This suggests that fold III may have lower aggregation potential than folds I and II. Note that there is currently no consensus that either haplotype leads to increased or decreased gene expression levels or TMEM106B protein abundance[2,5,8,10,14,15,17,25,26]. Another variant —rs1990620, which is in complete LD with the protective haplotype[10]—reportedly affects long-range chromatin interactions leading to decreased recruitment of CCCTC-binding factors (CTCF) decreasing

gene expression of *TMEM106B*[10]. However, multi-tissue eQTL analysis of rs1990622 using the UKBEC database shows no association[5], while a similar analysis of the protective rs1990622 and rs3173615 using the GTEx database indicates an association with *increased TMEM106B* expression in the cerebellum, cortex, and cerebellar hemisphere[5].

Thus far, studies have exclusively focused on single nucleotide variants from GWAS and/or short-read whole-genome sequencing data. Here, we explored whether a 330bp insertion, in linkage with the risk haplotype of *TMEM106B* may be the driver of the genetic association with altered disease risk.

Results

An AluYb8 element further differentiates the risk and protective haplotypes of TMEM106B

In a PacBio whole-genome long-read sequencing dataset of genomes from 92 Dutch Alzheimer Disease cases, 117 Dutch centenarians and 47 individuals from diverse human populations from HPRC (Human Pangenome Reference Consortium[27]), we observed a common ~330 bp insertion in the 3' untranslated region (UTR) in the risk haplotype of *TMEM106B* which is ~12.7 Kbp downstream of rs3173615. (see Table 1, see Figure 1). Annotations from the GRCh38 reference genome, along with sequence alignment to the DFAM repeat database[28], showed that the insertion is an Alu-mobile element of the Yb8 lineage sub-family. Alu-elements are retrotransposons that comprise ~11% of all human genomic sequences[29–31]. They propagate through a "copy-and-paste" mechanism by using

reverse-transcription of their mRNA transcripts to then insert themselves at different regions in the genome[29–31]. Importantly, Alu-elements are primate-specific; the Alu Yb8-lineage is among the youngest lineage of Alu-elements (integrated ~2.39 million years ago), and more dominantly expanded in humans than other primates[29,32,33].

Epigenetic differences between the TMEM106B haplotypes may yield functional consequences

Transcriptionally active mobile elements can dysregulate nearby gene expression and translation[30,34], hence, they are often supressed by the host through DNA methylation[35]. Therefore, we investigated potential differential methylation status of CpG islands between the *TMEM106B* haplotypes. Using the single-molecule methylation signals captured during PacBio HiFi whole-genome sequencing of the Dutch AD and centenarian genomes we identified a total of 231 unique CpG methylation sites in *TMEM106B* (see Figure 2A). Of these, a total of 14 CpG sites mapped in the AluYb8 element, most of which were methylated across all individuals, reflecting active suppression of retrotransposons[35](Fig 2C and 2D). However, methylation levels varied widely between individuals (Fig 2C and 2D), suggesting inter-individual variation in potential AluYb8 activation. Two specific CpG sites in the AluYb8 element had the highest variability: one within the "A box" RNA polymerase III recognition sequence[36,37]; and another in a GATA3 transcription factor recognition sequence[37]. Demethylation of such recognition sequences have been shown to increase transcription of Alu elements[33,38,39].

We found 39 sites for which methylation status was significantly different in those who carried at least one copy of the risk haplotype relative to non-carriers (Wilcoxon rank-sum

test, p-val $\leq 1.43x10^{-4}$); with 27 sites having at least a 10% higher methylation in carriers of the risk haplotype (see Figure 2C). Most of these sites affected intronic regions in *TMEM106B*, including other transposable elements from the L1-family (L1PA6, L1PA7, and L1ME3G), the J and S Alu-based sub-families (AluSg and AluJb), and the TcMar-Tigger family (Tigger1) (Figure 2B). Additionally, two sites affected two separate enhancer sequences: one immediately downstream of exon 1 (accession code EH38E2534496); and a second ~1,500 bp upstream of exon 2 (accession code EH38E2534502), which encodes for the 120 amino acid residues defining the C-terminus of TMEM106B protein found in the aggregated fibrils of neurodegenerative and aged human brains[20–24] (see Figure 2C). Notably, eight sites affected the 3' UTR of *TMEM106B* (outside the AlubY8 mobile element insertion), had at least a 10% higher methylation level in carriers of the risk haplotype compared to carriers of the protective haplotype (Figure 2B and 2C).

Only one of the 39 CpG sites was significantly *less* methylated in carriers of the risk haplotype compared to carriers of the protective haplotype. This site affected an MIRb mobile element—an ancient transposable element in mammalian genomes[40–42] (Figure 2C). Notably, MIRb sequences were found to be enriched in human long-noncoding RNA molecules (lncRNAs) that are localised in the nucleus[42], suggesting that these mobile elements also harbour gene and chromatin regulatory properties.

Together, next to differences in the genetic variation defining the protective and risk haplotypes of TMEM106B, we also observe marked differences between their epigenetic landscapes.

The 3' UTR of TARDBP is variably methylated in peripheral blood of AD and centenarian individuals

TARDBP encodes for the TDP-43 protein, which suppresses mobile element transcription but commonly aggregates in aged and neurodegenerative human brains [34,43,44]. The 3' UTR of *TARDBP* is reportedly demethylated with age, consequently dysregulating transcription and translation potentially reducing activity of TDP-43[45]. Hence, demethylation of *TARDBP* is thought to lead to increased Alu-element expression[35], which includes dysregulation of *TMEM106B*[30,34]. We thus investigated the methylation status of *TARDBP* in AD and centenarian genomes in context to the AluYb8 element insertion in the risk haplotype of *TMEM106B*.

Overall, we found 255 unique CpG sites across AD and centenarians, with an average of 240 total CpG sites per genome. We found no significant differences in CpG methylation per site between AD and centenarians (see Figure 3A). In the 3' UTR, we found 31 unique CpG sites, for which methylation status was highly variable between individuals, especially towards the end of the 3' UTR (see Figure 3B). This finding is in agreement with Koike *et al.* 2021, who reported that gene expression and alternative splicing of *TARDBP* can be dysregulated via age-related demethylation of a specific subregion in the 3' UTR harbouring a cluster of 15 CpG sites[45] (see black bar on top of Figure 3B). In particular, demethylation of the six left-most CpG sites were found to suppress alternative splicing and increase *TARDBP* gene expression[45] (Figure 4C). In our genomes, we identified 14/15 of these CpG sites: the six left-most sites showed similar demethylation patterns as those observed in controls and ALS patients from Koike *et al.* [45] (see Figure 3B and 3C). These findings suggest that age-related demethylation

of the 3' UTR of *TARDBP* also occurs in peripheral blood, which may lead to dysregulation of genes regulated by Alu-elements.

Long-read phasing of TMEM106B reveals novel haplotypes in West-African ancestry

The AluYb8 element is in complete LD with the reported variants characterising the risk haplotype of *TMEM106B* (r^2 = 0.98 – 1.0, $p < 2.2x10^{-16}$; Figure 1C) in Dutch genomes. Interestingly, LD is lower in the HPRC dataset (r^2=0.5-1.0, $p < 2.2x10^{-16}$, Figure 1C), which depended on a reduction in LD between rs13237518 (upstream from the AluYb8 element) and the SNPs downstream in the locus, which remain in complete linkage (Fig 1C). We found that this reduced LD was driven by 9 genomes with African ancestry: HG02630 (Gambian subpopulation), HG02717 (Gambian), HG02818 (Gambian), HG01891 (African Caribbean in Barbados), HG02055 (African Caribbean in Barbados), HG02723 (Gambian), HG03516 (Esan in Nigeria), NA18906 (Nigeria), and NA20129 (African Ancestry in Southwest USA). We hypothesised that this difference may be due to a (meiotic) recombination event(s) in the region between rs13237518 and the AluYb8, which may have given rise to an additional haplotype in the *TMEM106B* locus, unique for (Western-)Africans and their descendants.

To further characterize the different haplotypes in *TMEM106B,* we compared haplotype-phased *de novo* assemblies between the African genomes with reduced LD and the following European genomes: HG002 (Eastern European, Jewish-Ashkenazi ancestry), CHM13 (European ancestry[46]) genomes, and three Dutch centenarian genomes (Cent1, Cent2, and Cent3; Figure 4). We found a total of 341 unique variants across *TMEM106B* sequences, and

binary clustering revealed four major clusters. Two haplotype sequence clusters—termed groups 2 and 3—were composed mostly of West-African genomes. The AluYb8 element was absent in groups 1 and 3, and present in groups 2 and 4. Notably, group 3 harbours 1990622, rs1990621, and rs1990620 risk alleles, but not rs3173615 (T185S) and rs13237518. In contrast, group 2 only harbours the rs13237518 risk allele. Together, genomes from African ancestry carry alternate haplotype sequences in the in *TMEM106B* locus, defined by a breakpoint between rs13237518 and the AluYb8 element, which explains the observed reduced LD between the common variants that characterize the locus in individuals with European ancestry (see Figure 1C and 4).

Discussion

Long-read sequencing revealed that the risk haplotype of the *TMEM106B* locus includes an AluYb8 mobile element in the canonical transcript of the 3' UTR of *TMEM106B* gene. Alu-based insertions are known to (dys)regulate transcription and translation of nearby genes[30,47,48], making it a candidate driver of the neurodegenerative effect associated with the *TMEM106B* risk haplotype. We further observed that CpG islands in the risk and protective haplotypes were differentially methylated, potentially influencing *TMEM106B* transcription due to differential binding affinity of RNA polymerases and regulatory elements. TDP-43 has previously been shown to suppress mobile element transcription, including Alu-elements[34,43,44]. We found that the 3' UTR of *TARDBP*, where age-related demethylation reportedly dysregulates TDP-43[45], is variably demethylated in the peripheral blood of AD patients and in centenarians. Lastly, a comparison of complete haplotype sequences between

European and African genomes revealed four major haplotype groups in the *TMEM106B* locus, two of which are dominantly defined by African individuals, explaining why linkage disequilibrium (LD) of the risk alleles is reduced in *TMEM106B* in genomes with African ancestry.

Our results lead us to propose a mechanism in which the presence of the AluYb8 element in TMEM106B—along with age, TDP-43 and PGRN—leads to increased risk for neurodegeneration (Figure 5). Transcriptionally active Alu-elements can dysregulate transcription and translation of nearby genes[30,34], and propagate throughout the host genome, leading to genomic instability and DNA damage[35]. Hence, Alu-elements are often hypermethylated by their host to suppress activation[35]. Indeed, we found that the methylation level of the *TMEM106B* risk haplotype was higher than in the protective haplotype, likely a consequence of repression of the AluYb8 insertion[35]. However, aging commonly leads to a global decrease in methylation in humans[49,50], which may lead to increased expression of mobile elements and retrotransposition in (aged) humans[51–53]. Mobile elements are transcribed by RNA polymerase III[54], which is thought to compete with RNA polymerase II at overlapping (promoter) regions along the DNA[34,55]. This competition may ultimately lead to decreased gene expression and to decreased protein abundance of the nearby gene[34,55]. In line with this, methylation of the risk haplotype in AD and centenarians genomes was relatively reduced inside the binding site sequence of RNA polymerase III and nearby the binding sequence of GATA3, which may be indicative of age-related activation of the AluYb8, leading to dysregulation of *TMEM106B* gene transcription.

TMEM106B commonly co-aggregates with TDP-43 in various neurodegenerative diseases[56]. TDP-43 is encoded by the *TARDBP* gene across multiple cell-types and tissues, and is involved in RNA regulation, influencing transcription, mRNA stabilisation, and splicing[56,57]. Intriguingly, the 3' UTR of *TARDBP* is demethylated with age, which reportedly dysregulates *TARDBP* splicing and its expression levels[45]. In turn, this may contribute to the widely observed age-related TDP-43 aggregation, and loss of TDP-43 function[45]. Ultimately, this leads to cognitive impairments, as we recently observed that TDP-43 accumulation in the post-mortem brains of centenarians negatively correlates with cognitive function measured only a few months before death[58].

TDP-43 has been shown to regulate splicing, transcription and translation of nearby genes by suppressing Alu-element activation[34,43,44]. Age-related demethylation of the 3' UTR of *TARDBP* can suppress alternative splicing and increase gene expression, potentially leading to TDP-43 aggregation and decreased TDP-43 functionality[45]. Indeed, we observed similar patterns of demethylation in the 3' UTR of *TARDBP* as those observed in amyotrophic lateral sclerosis (ALS) and (healthy) control individuals[45] (see Figure 4). Notably, a recent study indicated that in iPSC-derived neurons, loss of TDP-43 function leads not only to increased mis-splicing of TDP-43 itself, but also of many other proteins including TMEM106B, likely by the loss of Alu-element suppression, leading to intron retention and the synthesis of alternate, non-functional, protein sequences.[57] Indeed, with the loss of TDP43 activity, expression levels of the AluYb8 element in the *TMEM106B* 3' UTR may be elevated, leading to the downregulation of *TMEM106B* expression.

Furthermore, the risk haplotype of *TMEM106B* associates more strongly with neurodegeneration in carriers of pathogenic variants in the *GRN* gene [2,3], which encodes the progranulin protein (PGRN). Both TMEM106B and PGRN proteins are essential for lysosome activity, and hence, impairment of both leads to profound additive, and possibly synergistic effects [12,13,59]. Interestingly, TDP-43 aggregation can be cleared by autophagy-lysosome pathway[60], which seems to be dependent on PGRN[12], as deficiency in PGRN leads to higher levels of TDP-43 aggregation[12,61,62]. As such, dysregulation of the lysosome by loss of PGRN may hamper TDP-43 degradation, leading to a stronger increase of TDP-43 aggregation[12]. Carriers of a pathogenic *GRN* variant, *and* the AluYb8 element in TMEM106B may be especially vulnerable for impaired lysosomal function. Vice versa, of all AD GWAS loci, the protective alleles in the *TMEM106B*, *SORT1*, and *GRN*—all involved in lysosome activity—are among the variants most strongly enriched in cognitively healthy centenarians[4], suggesting that maintaining endolysosomal function is highly important to maintain cognitive health until old age[4].

We further highlight that, even in a transcriptionally inactive state, the AluYb8 element in the 3' UTR can dysregulate *TMEM106B*. Impaired splicing, due to loss of TDP-43 is activity, may lead to the retainment of Alu-elements in mature mRNA transcript[30]. *TMEM106B* transcripts that include the AluYb8 element may be differentially suppressed by miRNAs, as Alu-elements are often micro-RNA (miRNA) targets[63,64], leading to mRNA degradation via Staufen-mediated decay[65]. Indeed, miRNAs have been shown to bind to the 3' UTR of *TMEM106B*, leading to transcriptional suppression[16]. Similarly, interactions with other RNA molecules, including Alu-elements or long non-coding RNAs (lncRNAs) in both *cis*- and *trans*-, may also lead to similar suppression and decay of *TMEM106B* transcripts[65]. Additionally, DNA-RNA complexes—

known as R-loops—can form during transcription[66], which can only be removed by specialised repair machineries requiring double-strand DNA break repair, potentially leading to DNA damage and genome instability[66]. Notably, reduced activity of TDP-43 in neuronal cell lines from an ALS patient showed increased accumulation of R-loop formation, likely due to loss of Alu elements-suppression[67].

Overall, the above interactions indicate that additional functional studies are needed to better understand how the Alu-element in the risk haplotype influences gene expression and/or translation of *TMEM106B.* We cannot rule out additional regulatory consequences of the presence/absence of the AluYb8 element, or possible epistatic effects of other variants on the haplotype, including the T185S substitution. Furthermore, we used DNA derived from peripheral blood as a resource for long-read sequencing of the genome; while this allows haplotyping, methylation status is tissue-specific, and should be further investigated in relevant (brain) tissues.[45] Nevertheless, we speculate that the presence or absence of the Alu-element may have affected experiments addressing TMEM106B function, leading to conflicting reports on transcriptional and translational effects of the risk and protective haplotype[2,5,8,10,14,15,17,25,26].

Next to the proposed mechanism of TMEM106B involvement in FTLD, this study also highlights how our limited understanding of neurodegenerative risk loci in non-Europeans may be driven by the lack of complete haplotype characterisations in other ancestral populations[68,69], along with limited sample size[68,69]. We observed four *TMEM106B* haplotype groups, differentially present in Europeans and African individuals. Based on the variational pattern of the four haplotype groups, we propose that the AluY8b element was inserted in

the ancestral haplotype 1, giving rise to haplotype 2; then haplotypes 2 and 3 may have merged through meiotic recombination, giving rise to haplotype 4. The availability of additional long read genomes from diverse population-backgrounds is necessary for a rigorous evaluation of the evolutionary development of TMEM106B haplotypes; we cannot exclude that additional *TMEM106B* haplotypes in the human population exist, nor that haplotype groups 2 and 3 are entirely unique to Africans. Our observations call for ancestry-specific imputation panels for individuals with non-European ancestry.

Conclusion

An AluYb8 mobile element residing in the 3'UTR of *TMEM106B* is commonly present in the human population, and is in complete linkage with the risk haplotype of *TMEM106B* in genomes from individuals with European descent. We particularly highlight that the effect of the Alu-element is enforced by specific states of PGRN, TDP-43, and most importantly, advanced age. We further propose that a negative feedback loop mediated by the Alu-element provides an interactive mechanism involving of TMEM106B, PGRN, TDP-43, and age (Figure 5). The Alu-element as the driving risk factor of the *TMEM106B* risk haplotype may allow the exploitation of existing evolutionary-selected defence mechanisms against retrotransposon transcription may provide strategies for development of therapeutics[41].

Methods

Sample populations

The 92 Alzheimer disease (AD) patients in this study are from the Amsterdam Dementia Cohort (ADC) and were diagnosed as previously described[70]. The 117 cognitively healthy

centenarians in this study are individuals from the 100-plus Study[71]. 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.

PacBio HiFi long-read whole-genome sequencing

Genomic DNA of AD and centenarian individuals were extracted from peripheral blood, and gDNA integrity was assessed with Qubit system 2 (Qubit dsDNA HS Assay Ref. No. Q32854) and Nanodrop (NanoDrop 2000). Specifically, we required DNA to have a purity of 260/230 of 2.00 and above, and 260/280 of 1.8 and above. When requirements were not met, a clean-up step was performed with 20 minutes incubation at room temperature with 1:1 AMPure PB beads (Pacific Biosciences Ref. No. 100-265-900) and two times washing with 80% ethanol. DNA is eluted with 120 µl Elution buffer (Pacific Biosciences Ref. No. 101-633-500). gDNA was sheared with the Megaruptor from Diagenode with a long hydropore cartridge (Cat. No. E07010003), using insert concentration of 116 µg with a shearing speed of first 30, and afterwards 31. Qubit dsDNA HS Assay (Ref. No. Q32854) was used to quantify DNA concentration. DNA damage repair was performed using SMRTbell Enzyme Clean Up Kit 2.0 (Ref. No. 101-932-600) with the reaction volumes as described in the manufacterers instructions. More specifically, a nuclease treatment on the formed SMRTbell library was done by preparing a master mix with SMRTbell Enzyme Clean Up Mix (Pacific Biosciences Ref. No.101-932-900), SMRTbell Enzyme Clean Up Buffer 2.0 (Pacific Biosciences Ref. No. 101-932-700) and Molecular biology Grade water. Reaction mix was added to the library complex and incubated at 37°C for 30 minutes. SMRTbell Library was purified using 1.x AMPurePB beads

(Pacific Biosciences Ref. No. 100-265-900) and two times washing with 80% ethanol, library complexes were eluted in 31 µl Elution buffer (Pacific Biosciences Ref. No. 101-633-500).

All libraries were prepared using either SMRTbell Template Prep Kit 2.0 (Pacific Biosciences Ref. No. 100-938-900) or 3.0 (Ref. no. 102-141-70) following manufacturer's instructions. Size selection was performed using the BluePippin high-pass DNA size selection (sage Science BLU0001; 0.75% Agarose Cassettes, with Marker S1). The range selection mode was set from 8,000-50,000 bp. Qubit dsDNA HS assay was used to quantify DNA concentration. Libraries having the desired size distributions were identified on the Femto pulse System (Agilent M5330AA) using the 48kb Ladder. Fractions centered between 15 kb, and 21 kb were used for sequencing.

Polymerase binding for all sequencing libraries were prepared using Sequel® II Binding Kit 2.2 (Pacific Biosciences Ref. No. 101-894-200) and 3.2 (Pacific Biosciences Ref. No. 102-194-100) following manufacturer's instructions. Sequencing primers were conditioned by heating to 80°C for 2 minutes and rapidly cooled to 4°C. The sequencing primer was annealed to the template at a molar ratio calculated by PacBio for 15 minutes at room temperature. After primer annealing, polymerase was bound to the primed template for 15 minutes at room temperature. Polymerase-bound samples were then kept at 4°C prior to use. Prior to sequencing, excess unbound polymerase was removed by incubating the complexes for 5 minutes with 1.2× (vol:vol) AMPure PB beads (Pacific Biosciences Ref. No. 100-265-900) at room temperature. Beads were not washed with 80% ethanol. After removing the free polymerase in the supernatant, the polymerase-bound complexes were eluted with Adaptive loading buffer (Pacific Biosciences Ref. No. 102-030-300). Sequencing reaction was performed

using the Sequel sequencing kit 2.0 (ref. no. 101-820-200) on a SMRT-cell 8M chip (ref. no. 101-389-001).

All samples were sequenced at the Department of Clinical Genetics at the Amsterdam University Medical Center (AUMC) using a Pacific Biosciences (PacBio) Sequel IIe System with 2 hours of pre-extension time and 30 hours of collection time. Overall, 92 AD genomes were sequenced at a median coverage of 18.2x, with median read-lengths of 14.8 Kbp; 117 centenarian genomes were sequenced at a median coverage of 20.1x, with median read-lengths of 14.7 Kbp.

PacBio HiFi data processing

After sequencing, raw reads were collected and analyzed through an in-house pipeline (freely available at https://github.com/holstegelab/snakemake_pipeline). Briefly, raw reads first underwent PacBio's *ccs* algorithm (v6.0.0, https://github.com/PacificBiosciences/ccs) to generate high-fidelity (HiFi) reads with custom parameters (min-passes 0, min-rq 0, keeping kinetics information). Using these parameters, we retained both *HiFi reads* (read-quality >99%, number of passes >3), and lower-quality *non-HiFi reads*, which are typically excluded when running the *ccs* algorithm with default settings. HiFi and non-HiFi reads were then aligned to GRCh38 (patch release 14) reference genomes using *pbmm2* (v1.3.0, https://github.com/PacificBiosciences/pbmm2), with suggested alignment presets for HiFi and non-HiFi reads, generating one BAM-file per SMRT-cell. When available, individual BAM-files from the same sample were merged into a single BAM-file using *samtools*[72] (v1.13).

We used HiFi-based publicly available whole-genomes to contextualise our findings to global human genetic diversity. Specifically, we downloaded 47 GRCh38-aligned BAM-files from the Human Pangenome Research Consortium (HPRC)[27] from their respective AWS S3 buckets representing "Year 1" data freeze (https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0). Long-reads were extracted from the BAM-files using *samtools* and re-aligned to GRCh38 using *minimap2*[73] (version 2.21-r1071).

Variant calling and linkage analysis

Structural variants across AD, centenarians, and HPRC genomes were identified using *sniffles2*[74] (version 2.0.6) with default parameters. Manual inspection of the structural variants in the TMEM106B gene using the UCSC genome browser (https://genome.ucsc.edu/) showed a large deletion overlapping the 3' UTR nearby known GWAS variants. The deletion exactly overlapped an AluYb8 mobile element based on annotations from the UCSC genome browser. This AluYb8-element (chr7:12242091-12242407) along with GWAS variants rs1020004, rs3173615, rs13237518, rs1990622, rs1990621, and rs1990620 were then locally assembled across all AD, centenarian, and HPRC genomes using the *"assemble"* command from *otter* (version 0.1.0) with the "*-F 0.2*" parameter. The locally assembled sequence of the deletion were aligned to DFAM[28] (release 3.7), confirming the structural variant is indeed an AluYb8 mobile element. The locally assembled sequences were then used to jointly genotype all variants using the "*genotype"* command from *otter* with the *"-e 0.5 -m 10*", ensuring that only the genotypes of variants with at least 10x of total coverage per region per sample were reported. Pairwise linkage-disequilibrium (LD) was then performed using the "genetics" package (version 1.3.8.1.3) in RStudio (version 2023.06.0+421). GWAS-variants and transcripts in TMEM106B were visualized with *ggbio*[75] (version 1.42.0) using the

EnsDB.Hsapiens.v86 package[76], as well as LocusZoom[77] using the "*Alzheimer_NatGenet2022*" study.

Comparison of complete *TMEM106B* haplotype sequences

Haplotype-phased *de novo* genome assembly was performed with *hifiasm*[78] (version 0.16.1-r375) using default parameters on three high-coverage (>30x) centenarian genomes based on PacBio HiFi whole-genome sequencing data generated in this study (Cent1, Cent2, and C3). Haplotype-phase *de novo* genome assemblies from the HPRC "Year 1 Data Freeze" dataset were downloaded[27] (see "PacBio HiFi data processing"). The centenarian genome assemblies, the CHM13 reference genome (version 2.0)[46], and HPRC genomes with reduced LD (see section above), along with HG002, were whole-genome aligned to GRCh38, and both single-nucleotide and (large) structural variation were identified with *paftools* (version 2.22-r1101, https://github.com/lh3/minimap2/blob/master/misc/README.md). Variant calls corresponding to the *TMEM106B* locus using GRCh38 coordinates (chr7:12211294-12244382) were merged, and their presence and absence across each haplotype sequence was binarised. Binary clustering was performed with *hclust* in RStudio[79].

Profiling methylation in *TMEM106B* locus

Methylation probabilities 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. CpG site coordinates were intersected against regulatory and mobile element annotations in GRCh38 from the UCSC genome browser. RNA polymerase III and transcription binding recognition

sequences within the AluYb8 element were cross-referenced against published annotations from Lu *et al.* 2014[37] and Conti *etl al.* 2015[36]. Statistical comparison between carriers of the risk and protective *TMEM106B* haplotypes were done using RStudio[79].

Acknowledgements

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. 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).

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.

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: NT, MH, SvdL, Bioinformatic analysis: AS.

References

1. Feng, T., Lacrampe, A. & Hu, F. Physiological and pathological functions of TMEM106B: a gene associated with brain aging and multiple brain disorders. *Acta Neuropathol. (Berl.)* **141**, 327–339 (2021).
2. Nicholson, A. M. & Rademakers, R. What we know about TMEM106B in neurodegeneration. *Acta Neuropathol. (Berl.)* **132**, 639–651 (2016).
3. Van Deerlin, V. M. *et al.* Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. *Nat. Genet.* **42**, 234–239 (2010).
4. Tesi, N. *et al. Cognitively Healthy Centenarians are genetically protected against Alzheimer's disease specifically in immune and endo-lysosomal systems*. http://medrxiv.org/lookup/doi/10.1101/2023.05.16.23290049 (2023) doi:10.1101/2023.05.16.23290049.
5. Hu, Y. *et al.* rs1990622 variant associates with Alzheimer's disease and regulates TMEM106B expression in human brain tissues. *BMC Med.* **19**, 11 (2021).
6. Tropea, T. F. *et al.* TMEM106B Effect on cognition in Parkinson disease and frontotemporal dementia. *Ann. Neurol.* **85**, 801–811 (2019).
7. Mao, F. *et al.* TMEM106B modifies TDP-43 pathology in human ALS brain and cell-based models of TDP-43 proteinopathy. *Acta Neuropathol. (Berl.)* **142**, 629–642 (2021).
8. Li, Z. *et al.* The TMEM106B FTLD-protective variant, rs1990621, is also associated with increased neuronal proportion. *Acta Neuropathol. (Berl.)* **139**, 45–61 (2020).
9. Van Der Zee, J. *et al.* TMEM106B is associated with frontotemporal lobar degeneration in a clinically diagnosed patient cohort. *Brain* **134**, 808–815 (2011).

10. Gallagher, M. D. *et al.* A Dementia-Associated Risk Variant near TMEM106B Alters Chromatin Architecture and Gene Expression. *Am. J. Hum. Genet.* **101**, 643–663 (2017).

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

12. Rhinn, H., Tatton, N., McCaughey, S., Kurnellas, M. & Rosenthal, A. Progranulin as a therapeutic target in neurodegenerative diseases. *Trends Pharmacol. Sci.* **43**, 641–652 (2022).

13. Butler, V. J. *et al.* Progranulin Stimulates the In Vitro Maturation of Pro-Cathepsin D at Acidic pH. *J. Mol. Biol.* **431**, 1038–1047 (2019).

14. Cabron, A.-S. *et al.* Lack of a protective effect of the Tmem106b "protective SNP" in the Grn knockout mouse model for frontotemporal lobar degeneration. *Acta Neuropathol. Commun.* **11**, 21 (2023).

15. Harding, S. R. *et al.* The TMEM106B risk allele is associated with lower cortical volumes in a clinically diagnosed frontotemporal dementia cohort. *J. Neurol. Neurosurg. Psychiatry* **88**, 997–998 (2017).

16. Chen-Plotkin, A. S. *et al. 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).

17. Nicholson, A. M. *et al.* TMEM106B p.T185S regulates TMEM106B protein levels: implications for frontotemporal dementia. *J. Neurochem.* **126**, 781–791 (2013).

18. T Vicente, C. *et al.* C-terminal TMEM106B fragments in human brain correlate with disease-associated *TMEM106B* haplotypes. *Brain* awad133 (2023) doi:10.1093/brain/awad133.

19. Lang, C. M. *et al.* Membrane Orientation and Subcellular Localization of Transmembrane Protein 106B (TMEM106B), a Major Risk Factor for Frontotemporal Lobar Degeneration. *J. Biol. Chem.* **287**, 19355–19365 (2012).

20. Schweighauser, M. *et al.* Age-dependent formation of TMEM106B amyloid filaments in human brains. *Nature* **605**, 310–314 (2022).

21. Jiang, Y. X. *et al.* Amyloid fibrils in FTLD-TDP are composed of TMEM106B and not TDP-43. *Nature* **605**, 304–309 (2022).

22. Chang, A. *et al.* Homotypic fibrillization of TMEM106B across diverse neurodegenerative diseases. *Cell* **185**, 1346-1355.e15 (2022).

23. Fan, Y. *et al.* Generic amyloid fibrillation of TMEM106B in patient with Parkinson's disease dementia and normal elders. *Cell Res.* **32**, 585–588 (2022).

24. Fan, Y. *et al.* Newly identified transmembrane protein 106B amyloid fibrils in the human brain: pathogens or by-products? *Ageing Neurodegener. Dis.* **3**, 4 (2023).

25. Ren, Y. *et al.* TMEM106B haplotypes have distinct gene expression patterns in aged brain. *Mol. Neurodegener.* **13**, 35 (2018).

26. Yang, H.-S. *et al.* Genetics of Gene Expression in the Aging Human Brain Reveal TDP-43 Proteinopathy Pathophysiology. *Neuron* **107**, 496-508.e6 (2020).

27. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).

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

29. Deininger, P. Alu elements: know the SINEs. *Genome Biol.* **12**, 236 (2011).

30. Chen, L.-L. & Yang, L. ALU ternative Regulation for Gene Expression. *Trends Cell Biol.* **27**, 480–490 (2017).

31. Solyom, S. & Kazazian, H. H. Mobile elements in the human genome: implications for disease. *Genome Med.* **4**, 12 (2012).

32. Feusier, J. *et al.* Discovery of rare, diagnostic AluYb8/9 elements in diverse human populations. *Mob. DNA* **8**, 9 (2017).

33. Carter, A. B. *et al.* Genome-wide analysis of the human Alu Yb-lineage. *Hum. Genomics* **1**, 167 (2004).

34. Morera, A. A., Ahmed, N. S. & Schwartz, J. C. TDP-43 regulates transcription at protein-coding genes and Alu retrotransposons. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1862**, 194434 (2019).

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

36. Conti, A. *et al.* Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Res.* **43**, 817–835 (2015).

37. Luo, Y., Lu, X. & Xie, H. Dynamic *Alu* Methylation during Normal Development, Aging, and Tumorigenesis. *BioMed Res. Int.* **2014**, 1–12 (2014).

38. Kochanek, S., Renz, D. & Doerfler, W. Transcriptional silencing of human Alu sequences and inhibition of protein binding in the box B regulatory elements by 5′-CG-3′ methylation. *FEBS Lett.* **360**, 115–120 (1995).

39. Jordà, M. *et al.* The epigenetic landscape of *Alu* repeats delineates the structural and functional genomic architecture of colon cancer cells. *Genome Res.* **27**, 118–132 (2017).

40. Smit, A. F. A. & Riggs, A. D. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.* **23**, 98–102 (1995).

41. Goodier, J. L. Restricting retrotransposons: a review. *Mob. DNA* **7**, 16 (2016).

42. Carlevaro-Fita, J. *et al.* Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. *Genome Res.* **29**, 208–222 (2019).

43. Li, W., Jin, Y., Prazak, L., Hammell, M. & Dubnau, J. Transposable Elements in TDP-43-Mediated Neurodegenerative Disorders. *PLoS ONE* **7**, e44099 (2012).

44. Tam, O. H. *et al.* Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Rep.* **29**, 1164-1177.e5 (2019).

45. Koike, Y. *et al.* Age-related demethylation of the TDP-43 autoregulatory region in the human motor cortex. *Commun. Biol.* **4**, 1107 (2021).

46. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

47. Deininger, P. L. & Batzer, M. A. Alu Repeats and Human Disease. *Mol. Genet. Metab.* **67**, 183–193 (1999).

48. Gussakovsky, D. & McKenna, S. A. Alu RNA and their roles in human disease states. *RNA Biol.* **18**, 574–585 (2021).

49. Ciccarone, F., Tagliatesta, S., Caiafa, P. & Zampieri, M. DNA methylation dynamics in aging: how far are we from understanding the mechanisms? *Mech. Ageing Dev.* **174**, 3–17 (2018).

50. Lee, J.-H., Kim, E. W., Croteau, D. L. & Bohr, V. A. Heterochromatin: an epigenetic point of view in aging. *Exp. Mol. Med.* **52**, 1466–1474 (2020).

51. Mustafina, O. E. The possible roles of human Alu elements in aging. *Front. Genet.* **4**, (2013).

52. Larsen, P. A. *et al.* The *Alu* neurodegeneration hypothesis: A primate-specific mechanism for neuronal transcription noise, mitochondrial dysfunction, and manifestation of neurodegenerative disease. *Alzheimers Dement.* **13**, 828–838 (2017).

53. De Cecco, M. *et al.* Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging* **5**, 867–883 (2013).

54. Zhang, X.-O., Gingeras, T. R. & Weng, Z. Genome-wide analysis of polymerase III–transcribed *Alu* elements suggests cell-type–specific enhancer function. *Genome Res.* **29**, 1402–1414 (2019).

55. Gao, Z., Herrera-Carrillo, E. & Berkhout, B. RNA Polymerase II Activity of Type 3 Pol III Promoters. *Mol. Ther. - Nucleic Acids* **12**, 135–145 (2018).

56. Jo, M. *et al.* The role of TDP-43 propagation in neurodegenerative diseases: integrating insights from clinical and experimental studies. *Exp. Mol. Med.* **52**, 1652–1662 (2020).

57. Seddighi, S. *et al. Mis-spliced transcripts generate* de novo *proteins in TDP-43-related ALS/FTD*. http://biorxiv.org/lookup/doi/10.1101/2023.01.23.525149 (2023) doi:10.1101/2023.01.23.525149.

58. Zhang, M. *et al.* The correlation between neuropathology levels and cognitive performance in centenarians. *Alzheimers Dement.* alz.13087 (2023) doi:10.1002/alz.13087.

59. Clayton, E. L. & Isaacs, A. M. Progranulin and TMEM106B: when two become wan. *EMBO Rep.* **21**, e51668 (2020).

60. Root, J., Merino, P., Nuckols, A., Johnson, M. & Kukar, T. Lysosome dysfunction as a cause of neurodegenerative diseases: Lessons from frontotemporal dementia and amyotrophic lateral sclerosis. *Neurobiol. Dis.* **154**, 105360 (2021).

61. Zhang, J. *et al.* Neurotoxic microglia promote TDP-43 proteinopathy in progranulin deficiency. *Nature* **588**, 459–465 (2020).

62. Beel, S. *et al.* Progranulin reduces insoluble TDP-43 levels, slows down axonal degeneration and prolongs survival in mutant TDP-43 mice. *Mol. Neurodegener.* **13**, 55 (2018).

63. Daskalova, E., Baev, V., Rusinov, V. & Minkov, I. 3'UTR-located *ALU* Elements: Donors of Potetial miRNA Target Sites and Mediators of Network miRNA-based Regulatory Interactions. *Evol. Bioinforma.* **2**, 117693430600200 (2006).

64. Pandey, R. *et al.* Alu-miRNA interactions modulate transcript isoform diversity in stress response and reveal signatures of positive selection. *Sci. Rep.* **6**, 32348 (2016).

65. Park, E. & Maquat, L. E. Staufen-mediated mRNA decay: Staufen-mediated mRNA decay. *Wiley Interdiscip. Rev. RNA* **4**, 423–435 (2013).

66. Niehrs, C. & Luke, B. Regulatory R-loops as facilitators of gene expression and genome stability. *Nat. Rev. Mol. Cell Biol.* **21**, 167–178 (2020).

67. Giannini, M. *et al.* TDP-43 mutations link Amyotrophic Lateral Sclerosis with R-loop homeostasis and R loop-mediated DNA damage. *PLOS Genet.* **16**, e1009260 (2020).

68. Lake, J. *et al.* Multi-ancestry meta-analysis and fine-mapping in Alzheimer's disease. *Mol. Psychiatry* (2023) doi:10.1038/s41380-023-02089-w.

69. Sherva, R. *et al.* African ancestry GWAS of dementia in a large military cohort identifies significant risk loci. *Mol. Psychiatry* **28**, 1293–1302 (2023).

70. Van Der Flier, W. M. & Scheltens, P. Amsterdam Dementia Cohort: Performing Research to Optimize Care. *J. Alzheimers Dis.* **62**, 1091–1111 (2018).

71. Holstege, H. *et al.* The 100-plus Study of cognitively healthy centenarians: rationale, design and cohort description. *Eur. J. Epidemiol.* **33**, 1229–1249 (2018).

72. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

73. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

74. Smolka, M. *et al. Comprehensive Structural Variant Detection: From Mosaic to Population-Level*. http://biorxiv.org/lookup/doi/10.1101/2022.04.04.487055 (2022) doi:10.1101/2022.04.04.487055.

75. Yin, T., Cook, D. & Lawrence, M. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.* **13**, R77 (2012).

76. Rainer, J. EnsDb.Hsapiens.v86. (2017) doi:10.18129/B9.BIOC.ENSDB.HSAPIENS.V86.

77. Boughton, A. P. *et al.* LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics* **37**, 3017–3018 (2021).

78. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

79. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2021).

Table 1. An AluYb8 mobile element is commonly present in the 3' UTR of TMEM106B. Observed relative frequencies of homozygous non-carriers (0/0), heterozygous (0/1), and homozygous carriers (1/1) of the AluYb8 mobile elements in long-read sequencing datasets of Alzheimer's disease (AD) and cognitively healthy centenarian genomes from The Netherlands (100-plus study), as well as genetically diverse genomes from the Human Pangenome Research Consortium (HPRC).

Study Cohort	phenotype	Genomes	0/0	0/1	1/1
ADC	AD patients	88	17% (15)	44% (39)	39% (34)
100-plus	centenarians	112	27.5% (31)	37.5% (42)	35% (39)
HPRC	Normal controls	47	17% (8)	43% (20)	40% (19)

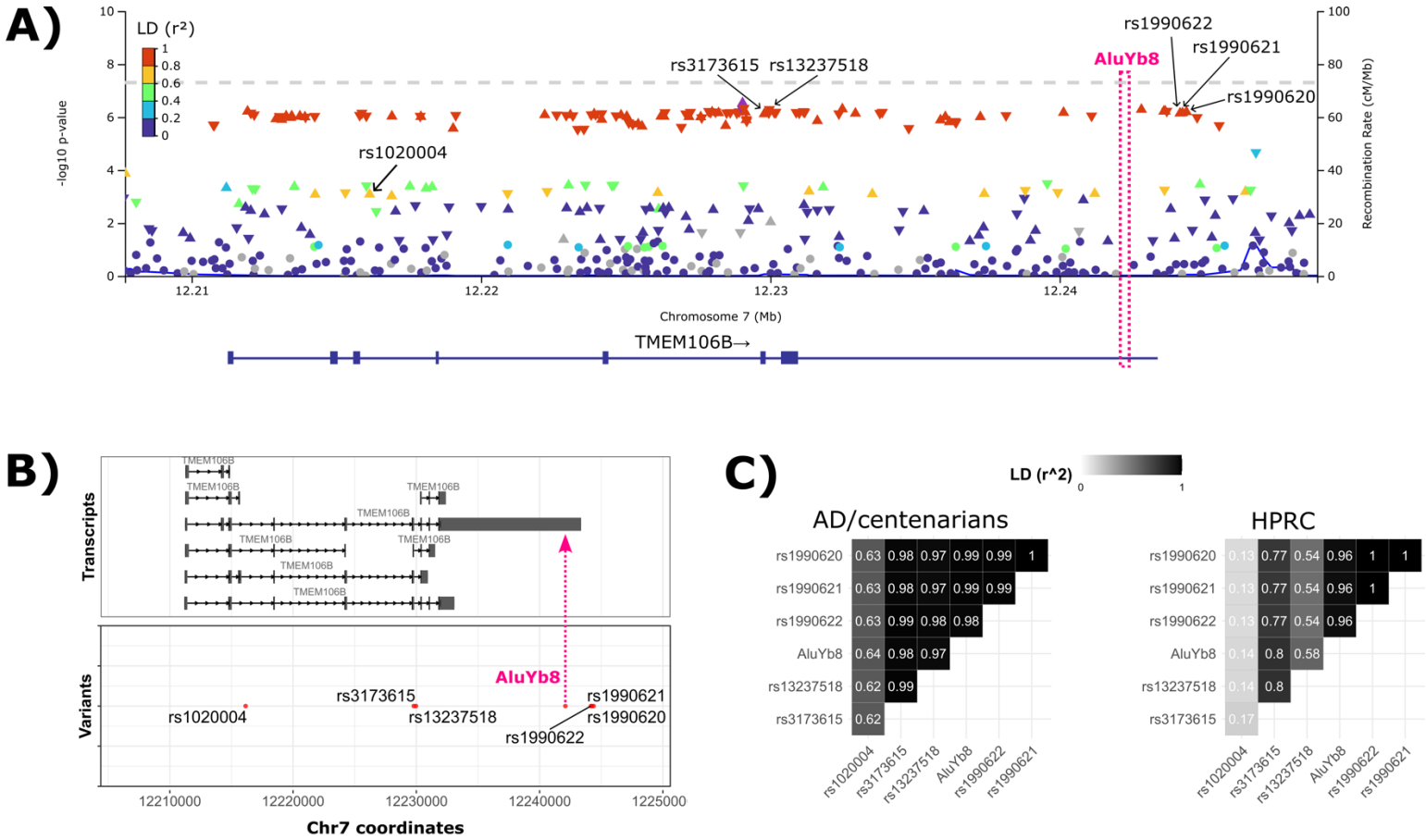

Figure 1. An AluYb8 mobile element further characterizes risk and protective haplotypes of *TMEM106B* associated in neurodegeneration. (A) Modified overview of the TMEM106B gene from LocousZoom[77] displaying genome-wide associated (GWAS) variants (triangles) from the Bellenguez *et al.* 2022 study[11]; the color of each variant represents LD (r^2) values based on genomes with European ancestry. **(B)** Annotated transcripts of TMEM106B from Ensembl[76] (top), along with the corresponding location of the AluYb8-element and known GWAS-variants (bottom). **(C)** LD (r^2) values based on 92 AD genomes and 117 centenarian genomes from The Netherlands[4,71] (left), and 43 genomes from Human Pangenome Reference Consortium (HPRC)[27] (right).

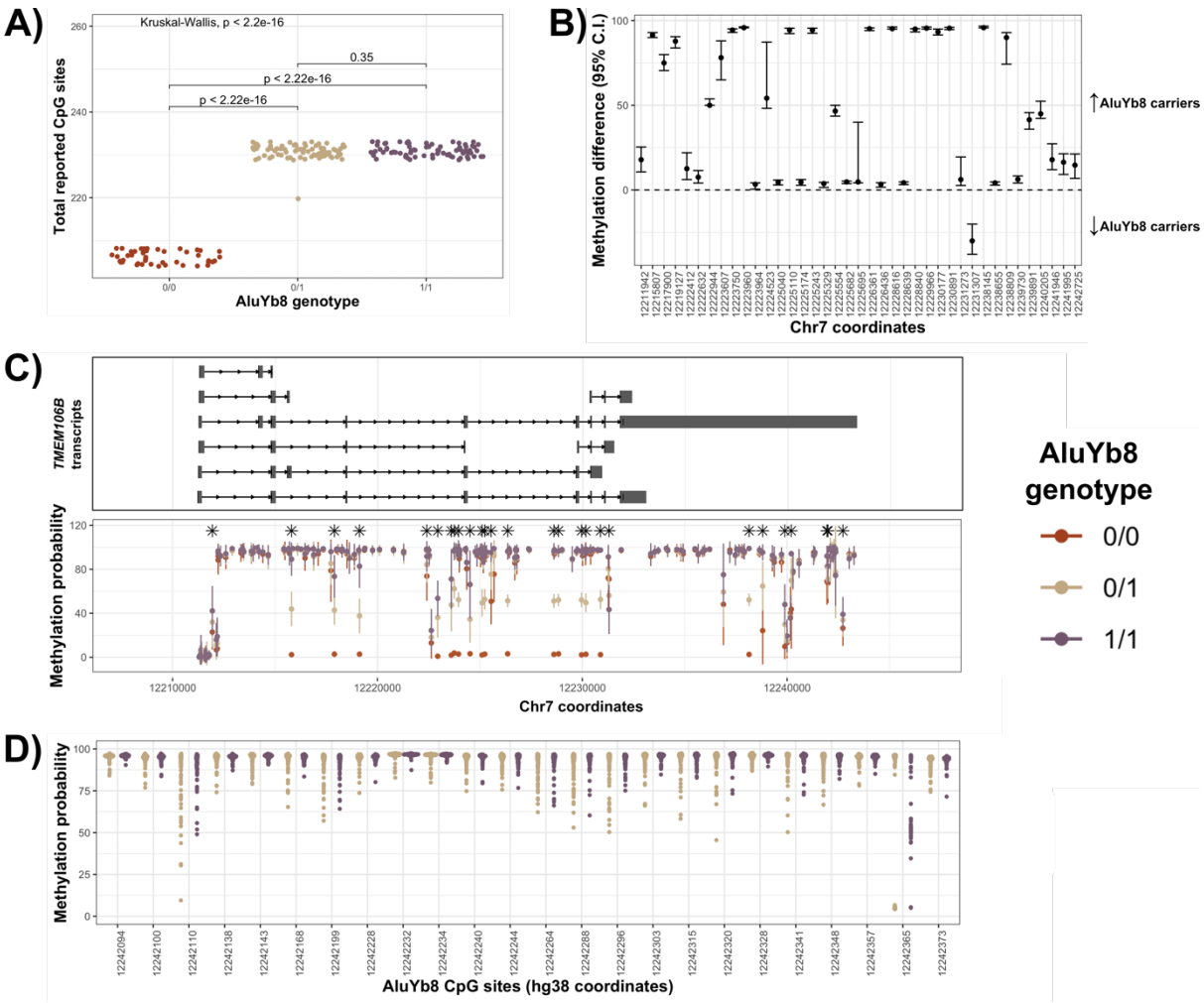

Figure 2. The risk and protective haplotype of *TMEM106B* vary in their epigenetic landscapes. Distribution of methylated CpG sites in *TMEM106B* based on PacBio HiFi whole-genome sequencing of peripheral blood from AD and centenarians. **(A)** Total number of reported CpG sites (Y-axis), partitioned by carriers of the AluYb8 mobile element in the 3' UTR of *TMEM106B* (X-axis), where 0 and 1 indicates non-carriers and carriers, respectively. Note that the risk haplotype is in complete linkage disequilibrium with the AluYb8 element (see Figure 1C). **(B)** 38 CpG sites were found to be significantly different between homozygous carriers (1/1) and homozygous non-carriers (0/0) (Wilcoxon rank-sum test, p-value $\leq 2.42\times10^{-4}$, Bonferroni corrected). Y-axis shows the estimated difference with the 95% confidence interval, while X-axis shows the coordinates of the CpG sites in the GRCh38 reference genome. CpG sites in carriers are generally more methylated. **(C)** *TMEM106B* transcripts from *Ensembl*[76] (top), overlayed with relative coordinates of each CpG site (bottom), displaying the average methylation probability (dots) along with the corresponding standard deviations (vertical bars). Asterisks indicate significantly different CpG sites from (B). **(D)** Methylation probability distributions (Y-axis) for each CpG site inside the AluYb8 element (X-axis) between heterozygous and homozygous AluYb8 carriers.

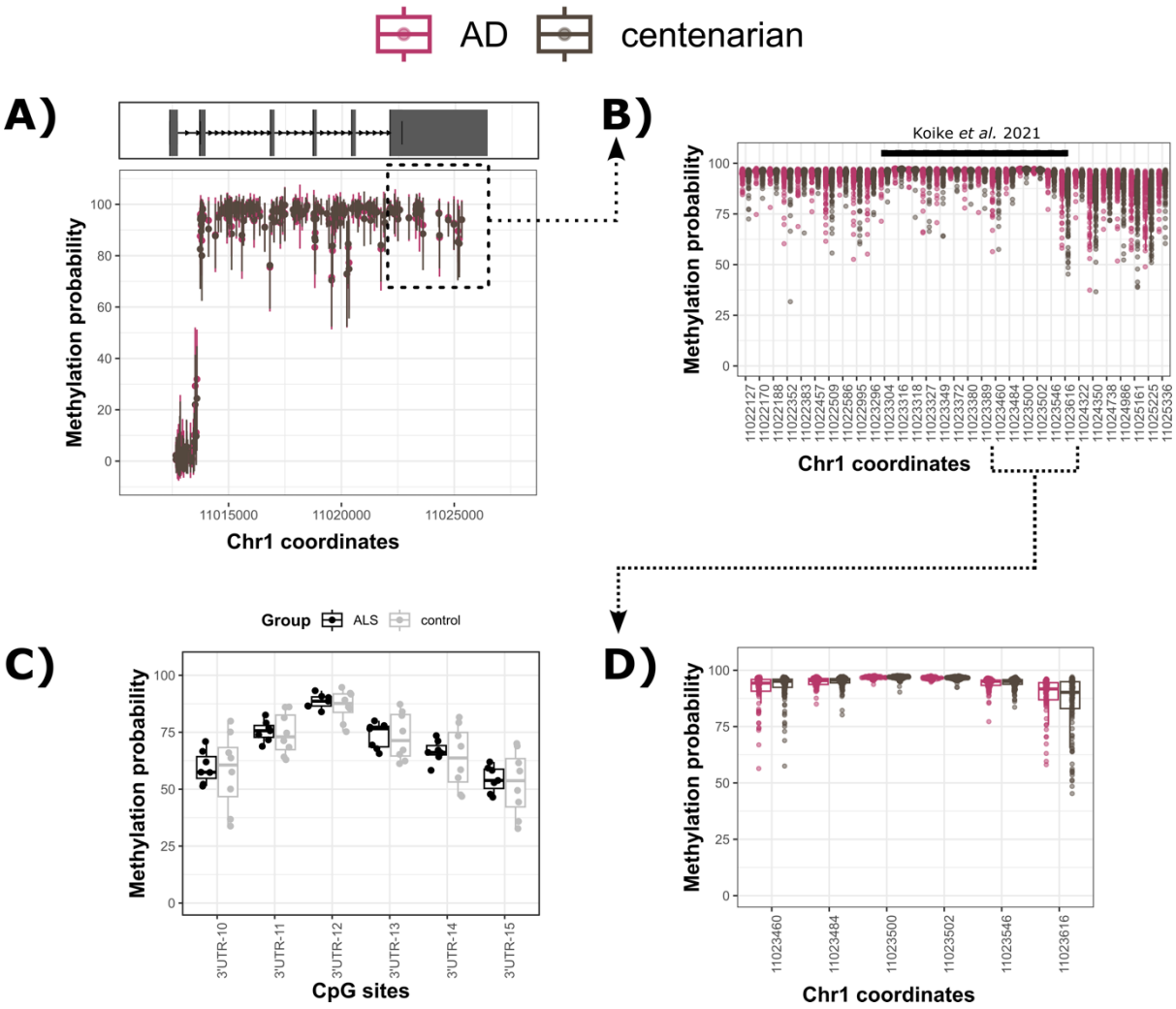

Figure 3. The 3' UTR of *TARDBP* is variably methylated. Distribution of methylated CpG sites in *TARDBP* based on PacBio HiFi whole-genome sequencing of peripheral blood from AD and centenarians. **(A)** Canonical transcript of *TARDBP* based on *Ensembl*[76] annotations (top), with the corresponding methylation status of CpG sites found within *TARDBP* region between AD (violet) and centenarian (dark grey) individuals (bottom). **(B)** Same as (A), but displaying only CpG sites within the 3' UTR of *TARDBP*. The black horizontal bar on top depicts CpG cluster in *TARDBP* reported by Koike *et al*. 2021[45], where demethylation of **(C, D)** the six left-most CpG sites can dysregulate *TARDBP* transcription. **(C)** Displays methylation data for the six left-most CpG sites from motor neurons from amyotrophic lateral sclerosis (ALS) patients and control individuals as reported by Koike *et al*. 2021[45]. **(D)** Same as (C), but from peripheral blood from AD and centenarian individuals from this study

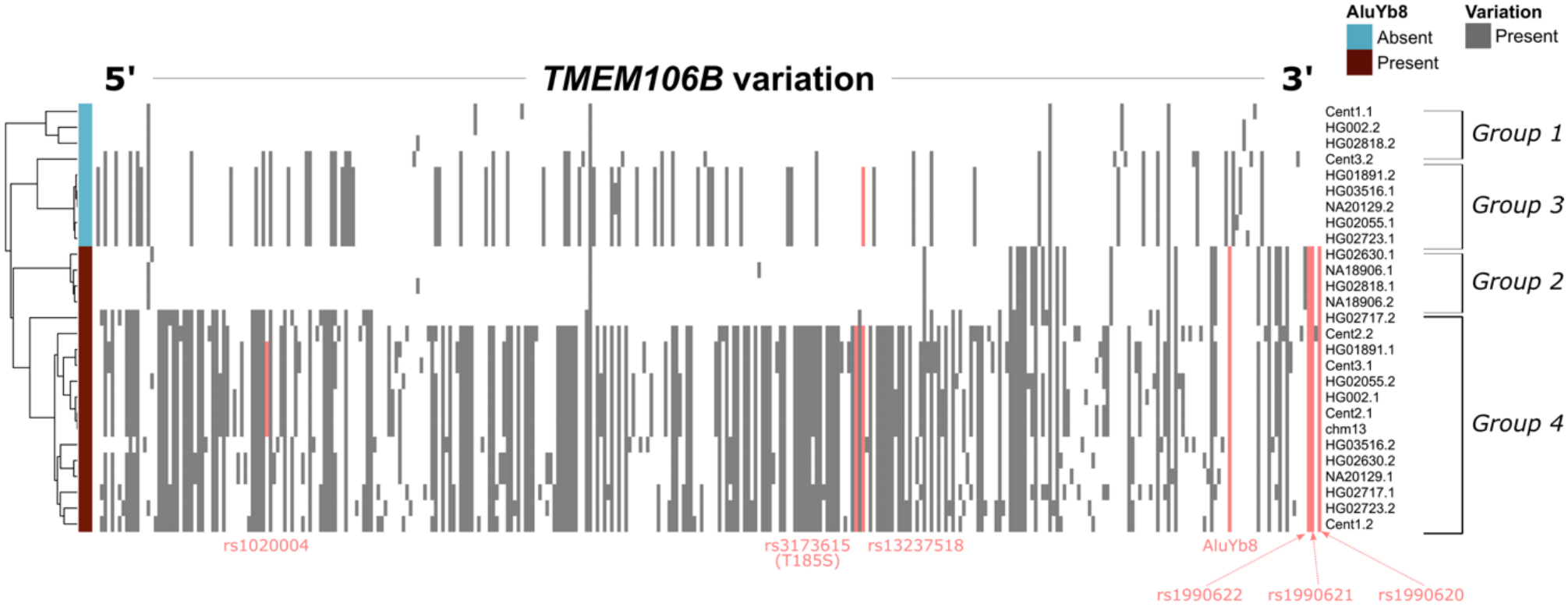

Figure 4. Comparison of phased *de novo* assemblies reveals novel haplotypes in *TMEM106B* in genomes with African ancestry.
Single-nucleotide and (large) structural variants were identified in 27 complete phased sequences of *TMEM106B* from 14 long-read *de novo* genome assemblies, where CHM13, HG002, and Cent1-3 are from European ancestry; the rest are genomes from African ancestry with reduced linkage disequilibrium (see Figure 1). Each row represents one of two haplotype sequences (.1 or .2 suffix) coloured by presence (maroon) or absence (blue) of the AluYb8 element in the 3' UTR. Columns represent all variation jointly identified in all *TMEM106B* haplotype sequences, ordered from their relative position from 5' to 3' orientation; grey bars represent the binary presence of each variant in each haplotype sequence. Selected variants previously associated with neurodegeneration (see Figure 1) are highlighted (salmon colour) along with the AluYb8 element. Clustering of the binary presence/absence of all variation yields four major clusters, annotated as haplotype groups 1 to 4

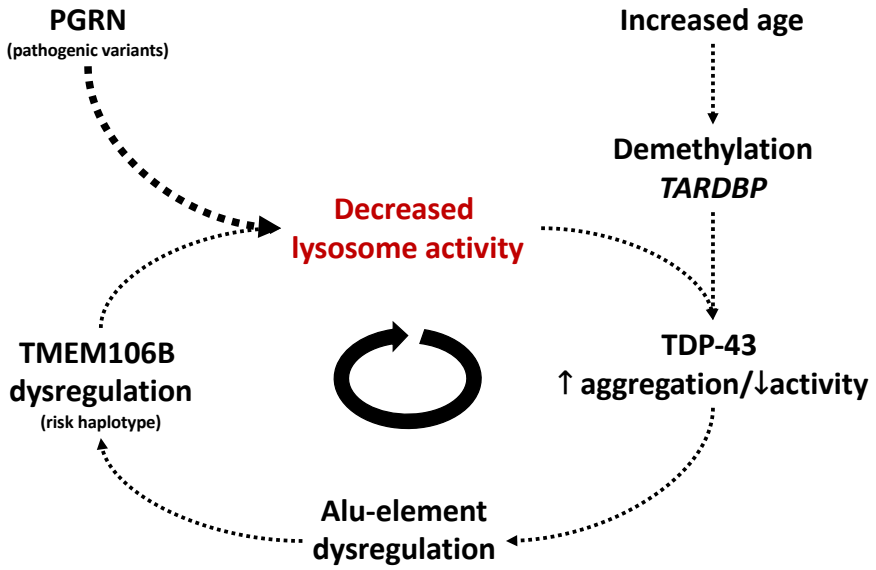

Figure 5. Endolysosomal processes depend on GRN, TMEM106B, TDP-43 and advanced age interactions. Pathogenic *GRN* variants leads to compromised endolysosomal activity, in turn increasing aggregation of TDP-43 due to incomplete degradation. Increased age leads to hypomethylation of *TARDP* (gene encoding TDP-43), which dysregulates TDP-43 protein synthesis. As TDP-43 supresses Alu-element transcription, reduced functionality of TDP-43 leads to increased Alu-element transcription, consequently dysregulating TMEM106B and further compromising endolysosome activity.