A 3' UTR Deletion Is a Leading Candidate Causal Variant at the *TMEM106B* Locus Reducing Risk for FTLD-TDP

Augustine Chemparathy[1,2], Yann Le Guen[2], Yi Zeng[3], John Gorzynski[3,4], Tanner Jensen[3], Nandita Kasireddy[2], Lia Talozzi[2], Michael E. Belloy[2], Ilaria Stewart[2], Aaron D. Gitler[3], Anthony D. Wagner[5,6], Elizabeth Mormino[5,6], Victor W. Henderson[2,7], Tony Wyss-Coray[2], Euan Ashley[3,4], Michael D. Greicius[2]

[1]Stanford University School of Medicine, Stanford, CA
[2]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA
[3]Department of Genetics, Stanford University School of Medicine, Stanford, CA
[4]Division of Cardiology, Department of Medicine, Stanford University School of Medicine, Stanford, CA
[5]Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, USA
[6]Department of Psychology, Stanford University, Stanford, CA, USA.
[7]Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA.

Corresponding author:
Michael D. Greicius
Stanford Neuroscience Health Center
290 Jane Stanford Way, Rm E261
Stanford, CA 94305-5090
greicius@stanford.edu
(650) 498-4624

Abstract

Single nucleotide variants (SNVs) near *TMEM106B* have been associated with risk of frontotemporal lobar dementia with TDP pathology (FTLD-TDP) but the causal variant at this locus has not yet been isolated. The initial leading FTLD-TDP genome-wide association study (GWAS) hit at this locus, rs1990622, is intergenic and is in linkage disequilibrium (LD) with a *TMEM106B* coding SNV, rs3173615. We developed a long-read sequencing (LRS) dataset of 407 individuals in order to identify structural variants associated with neurodegenerative disorders. We identified a prevalent 322 base pair deletion on the *TMEM106B* 3' untranslated region (UTR) that was in perfect linkage with rs1990622 and near-perfect linkage with rs3173615 (genotype discordance in two of 274 individuals who had LRS and short-read next-generation sequencing). In Alzheimer's Disease Sequencing Project (ADSP) participants, this deletion was in greater LD with rs1990622 (R^2=0.920916, D'=0.963472) than with rs3173615 (R^2=0.883776, D'=0.963575). rs1990622 and rs3173615 are less closely linked (R^2=0.7403, D'=0.9915) in African populations. Among African ancestry individuals in the ADSP, the deletion is in even greater LD with rs1990622 (R^2=0.936841, D'=0.976782) than with rs3173615 (R^2=0.764242, D'=0.974406). Querying publicly available genetic datasets with associated mRNA expression and protein levels, we confirmed that rs1990622 is consistently a protein quantitative trait locus but not an expression quantitative trait locus, consistent with a causal variant present on the *TMEM106B* 3'UTR. In summary, the *TMEM106B* 3' UTR deletion is a large genetic variant on the *TMEM106B* transcript that is in higher LD with the leading GWAS hit rs1990622 than rs3173615 and may mediate the protective effect of this locus in neurodegenerative disease.

Introduction

Next-generation sequencing (NGS) has enabled genome-wide identification of single nucleotide variants (SNVs) associated with heritable diseases ranging from cancer to neurodegeneration. Variants identified as disease risk-modulating in genome-wide association studies (GWAS) of NGS data are frequently intergenic or intronic; such GWAS hits may be the causal variant themselves or may be in linkage with a nearby genetic feature that is the true disease risk-modifying variant.

Structural variants (SVs) – which include large insertions, deletions, duplications, and other genomic features greater than 50 base pair (bp) in length – are a source of genetic diversity whose impact on protein function is often readily interpretable due to their large size. SVs are presently challenging to identify with NGS. Whereas the 150 bp read length used by most NGS approaches is able to obtain high coverage of SNVs and some small insertions and deletions, SVs that exceed this read length are principally detected in NGS data by analyzing paired and split-read evidence as well as changes in sequencing depth[1–3]. Emerging long-read sequencing (LRS) technology utilizes reads of greater than 10 kilobases (kb), enabling large SVs to be directly sequenced and properly aligned to the genome[4]. LRS greatly improves on SV discovery over short-read NGS approaches and is able to identify more than twice as many SVs as ensemble methods operating on short-read NGS data; up to 83% of insertions identified by LRS are not detected by NGS algorithms[5].

We carried out LRS and SV calling for participants enrolled in the Stanford Alzheimer's Disease Research Center (ADRC) and/or the Stanford Aging and Memory Study (SAMS)[6]. After filtering for variants overlapping genes involved in Alzheimer's Disease (AD) and neurological function, we identified a highly prevalent SV on *TMEM106B*. A set of SNVs on and near *TMEM106B* have been associated with a strong protective effect in neurodegenerative diseases including

frontotemporal lobar dementia with TAR DNA-binding protein pathology (FTLD-TDP) and Alzheimer's Disease (AD) (though the effect size in AD is considerably smaller)[7]. We pursued the possibility that SVs on *TMEM106B* may mediate the protective effect of this locus.

The protective association of *TMEM106B* SNVs in FTLD-TDP was first described in a 2010 GWAS of FTLD-TDP cases and controls that identified three significant SNVs (rs1990622, rs6966915, rs1020004) in high linkage disequilibrium (LD) with one another, all on or near *TMEM106B*[8]. For the leading SNV, rs1990622, the major allele (rs1990622_A) was risk-increasing in FTLD-TDP (odds ratio (OR)=1.64) while the minor allele (rs1990622_G) was protective (OR=0.61). The minor allele of rs1990622 is quite common across populations in gnomAD[9] (allele frequency (AF)=0.4989) and has been associated with increased plasma levels of progranulin (GRN) in controls, suggesting that *TMEM106B* variants may protect against FTLD-TDP by rescuing *GRN* haploinsufficiency[10,11]. In the latest, largest AD GWAS a *TMEM106B* SNV (rs13237518, chr7:12229967, minor/major allele A/C) is found to be protective (OR=0.96)[7], though less so than rs1990622 in FTLD-TDP. In LDpair[12], rs13237518 is in LD with rs1990622 (R^2= 0.8779, D'=0.9409).

The mechanism by which variants at the *TMEM106B* locus affect FTLD-TDP risk remains unclear. The genome-wide significant *TMEM106B* SNVs identified in Van Deerlin et al. are either intronic or intergenic. The only coding SNV in high LD with rs1990622 is the missense variant rs3173615, which results in a p.T185S amino acid change in exon 6 of *TMEM106B*. An *in vitro* study found that overexpression of TMEM106B S185 in HeLa cells resulted in a smaller increase in TMEM106B protein levels than overexpression of WT TMEM106B, indicating that rs3173615 may hasten protein degradation[13]. However, an *in vivo* study using a *GRN*$^{-/-}$ mouse model homozygous for TMEM106B S186 (the conserved residue in mice) did not observe a change in TMEM106B protein levels relative to wild-type, nor amelioration of microgliosis or the

pathological lysosomal phenotype[14]. In addition, an FTLD-TDP GWAS evaluating both rs1990622 and rs3173615 found that rs1990622 was the most significant SNV at the *TMEM106B* locus following meta-analysis[15]. Thus, the evidence for rs3173615 as the causative variant on *TMEM106B* remains mixed.

In this study, we developed and queried a large LRS dataset and identified an SV on *TMEM106B* that may be the protective variant at the *TMEM106B* locus in FTLD and AD.

Methods

Study participants

Study protocols were approved by the Stanford University Institutional Review Board. The Stanford Alzheimer's Disease Research Center (ADRC) is a cohort of healthy older controls and patients with AD and related neurological disorders (n=274 with LRS and NGS, ages 36-93, 145 F, 129 M, healthy controls = 133, mild cognitive impairment cases = 47, AD cases = 22, other diagnosis = 72). All participants underwent a history and neurological exam, cognitive testing, and blood draw. Most participants also underwent brain imaging including MRI and amyloid PET scanning. Roughly 1/3 of participants also provided cerebrospinal fluid (CSF). Diagnoses were determined in a consensus conference meeting comprised of neurologists and neuropsychologists using standard clinical criteria for AD, MCI, and related disorders such as Parkinson's disease and Lewy body disease.

The Stanford Aging and Memory Study (SAMS) is a cohort of cognitively unimpaired older individuals (n=133 with LRS, ages 60-88, 73 F, 60 M). SAMS eligibility criteria include normal or corrected vision and hearing, native English speaking, no neurologic or psychiatric disease history, Clinical Dementia Rating score of zero, and normal performance on standardized

neuropsychological testing. Participants underwent CSF and plasma collection and brain imaging including MRI and amyloid PET. Unimpaired cognitive status was confirmed in a consensus conference meeting comprised of neurologists and neuropsychologists using standard clinical criteria.

Long-read sequencing, alignment and SV calling

High molecular weight DNA was extracted from primary blood mononuclear cells (PBMC's) that had been stored at -80C using a Puregene kit (Qiagen, Germany). DNA was sheared using a G-tube (Covaris LLC, Massachusetts). Sequencing libraries were prepared using Nanopore LSK-110 and sequenced on the PromethION48 (Oxford Nanopore Technologies, United Kingdom). An average of 50.4 gigabases were sequenced per sample, with a read length N50 of 18 kb. Sequencing data were base called using Guppy (High Accuracy, version 6.3), and aligned to HG38 using Minimap2[16]. Structural variants were called using Sniffles2[17] in population mode. Variants with start position overlapping *TMEM106B* were extracted.

Short-read next-generation sequencing

TMEM106B SNV genotypes were determined from short-read NGS performed at either the Beijing Genomics Institute (BGI) in Shenzhen, China or as part of the Stanford Extreme Phenotypes in Alzheimer's Disease project with sequencing performed at the Uniformed Services University of the Health Sciences (USUHS) on an Illumina HiSeq platform. Among the 274 ADRC participants, 29 participants were sequenced via USUHS and 245 via BGI. The Genome Analysis Toolkit (GATK) workflow Germline short variant discovery was used to map genome sequencing data to the reference genome (GRCh38) and to produce high-confidence variant calls using joint-calling[18].

3' UTR deletion dose curation

The genotypes of rs1990622, rs3173615, and the 3' UTR deletion were extracted for participants for whom whole genome LRS and NGS were available. Participants with discordant doses of the three variants in LRS – where the dose of any of the three variants differed from any other – were identified for manual curation. For these participants, LRS genome alignments were visualized in IGV[19] and the dose of the *TMEM106B* 3' UTR deletion was determined by the following criteria: (1) the dose was set to 0 if no reads contained the deletion, (2) the dose was set to 1 if at least one but not all reads contained the deletion, and (3) the dose was set to 2 if all reads contained the deletion.

Alzheimer's Disease Sequencing Project (ADSP) LD analysis

ADSP R3 SNVs and Biograph[20] SV calls were downloaded from NIAGADS (https://dss.niagads.org/datasets/ng00067/#data-releases). SNVs were subset to rs3173615 (7:12229791:C:G) and rs1990622 (7:12244161:A:G) using Plink 1.9[21]. The *TMEM106B* 3' UTR deletion (chr7:12242077; SVLEN=-322; SVTYPE=DEL) was identified in Biograph SV calls. Samples present in both SV and SNV data were identified, VCF files for both datasets were subset to these samples, and the files were concatenated using bcftools[22]. LD was computed using Plink 1.9 with the --r2 dprime flag.

To identify African ancestry individuals in the ADSP, ancestries of all ADSP individuals were determined using SNPWeights v2[23] using reference populations from the 1000 Genomes Consortium[24]. Individuals with greater than 75% African ancestry were classified as African ancestry.

eQTL and pQTL analysis

All expression quantitative trait locus (eQTL) effect sizes and p-values were queried for rs1990622 from summary statistics (Sieberts meta-analysis[25], CommonMind Consortium[26],

GTEx[27], Wingo[28], MetaBrain[29], eQTLGen[30]). Protein quantitative trait locus (pQTL) effect sizes and p-values for ARIC[31], DECODE[32], Wingo, and Banner[33] were also queried for rs1990622 from summary statistics. See Data Availability Statement for direct links to summary statistics queried. For ROSMAP brain areas BA9, BA6, and BA37, processed TMT quantitated protein abundance data from Synapse projects syn25006657 and syn2580853 were used to calculate effect size and p-values using a multiple linear regression in R. The *lm* function was used to fit a linear regression model to combine AMP-AD WGS and SNP array data for rs1990622 against proteomics data, covarying out the first three genetic principal components, *APOE* status, and diagnosis. The same linear model was computed for MSBB BA36 using processed TMT proteomics data from Synapse project syn25006647. See Data Availability Statement for direct links to raw protein abundance data used from Synapse.

Results

We carried out whole genome LRS and SV calling for 407 participants enrolled in the Stanford ADRC and/or SAMS. Two unique SVs overlapping *TMEM106B* were identified, as summarized in Figure 1a. One SV, a 322 base pair (bp) deletion located in the 3' untranslated region (UTR) of *TMEM106B*, is highly prevalent with AF=0.4568 in our LRS dataset, comparable to the AF in gnomAD of rs1990622 (AF=0.4989) and rs3173615 (AF=0.4902). The *TMEM106B* 3' UTR deletion was detected in both LRS and NGS (Figure 1b). The second SV was much less prevalent (AF=0.1096) than rs1990622 and rs3173615. We linked LRS data to high coverage NGS data for 274 Stanford ADRC participants in order to evaluate the LD between the 3' UTR deletion, rs1990622 and rs3173615. The *TMEM106B* 3' UTR deletion was in perfect concordance with rs1990622 and was concordant with rs3173615 in all but two individuals.

The LD between the *TMEM106B* 3' UTR deletion, rs1990622, and rs3173615 was established in a large cohort by querying the ADSP database. 16882 samples were genotyped at both

SNVs rs1990622 and rs3173615. The *TMEM106B* 3' UTR deletion was identified in 12120 of 16841 samples (AF=0.4977) in the Biograph SV callset provided by ADSP. The SV was in greater LD with rs1990622 (R^2=0.920916, D'=0.963472) than with rs3173615 (R^2=0.883776, D'=0.963575). In LDpair, rs1990622 and rs3173615 are in high LD when assessed across all populations (R^2=0.91, D'=0.9905). However, we noted that these SNVs are not as closely linked in LDpair in African populations (R^2=0.7403, D'=0.9915). In African ancestry individuals in ADSP, the *TMEM106B* 3' UTR deletion was in even greater LD with rs1990622 (R^2=0.936841, D'=0.976782) than with rs3173615 (R^2=0.764242, D'=0.974406). Taken together, these data indicate that the *TMEM106B* 3' UTR deletion is more closely associated with rs1990622 than rs3173615 across populations and may underlie the slightly greater significance of rs1990622 over rs3173615 in reducing risk of FTLD-TDP.

We evaluated the effect of rs1990622 on *TMEM106B* expression and protein levels in eQTL and pQTL datasets (Table 1). rs1990622 does not result in a significant effect on *TMEM106B* expression levels in seven datasets and has a significant effect in three datasets. Across eQTL datasets, the direction of the effect of rs1990622 varies ($\beta<0$ in two datasets, $\beta>0$ in five datasets). rs1990622 results in a statistically significant effect on TMEM106B protein levels in ten datasets and does not have a significant effect in one dataset (Wingo meta-analysis, p=0.1005). The effect size is negative in eight pQTL datasets and positive in three. Our observation of significant pQTL effects in the absence of a consistent eQTL effect is most consistent with a model in which the *TMEM106B* causative variant exerts its protective effect after transcription of *TMEM106B*.

Discussion

The *TMEM106B* 3' UTR deletion is a previously unreported variant that is a potential candidate to mediate the protective effect of the *TMEM106B* locus in FTLD-TDP. At present, the

candidates for the causal variant at this locus are (1) the *TMEM106B* 3' UTR deletion; (2) rs3173615, the only coding SNV in this linkage block; (3) rs1990622, the leading GWAS hit; or (4) another variant in this linkage block. Recent *in vivo* work using a mouse model demonstrated that homozygous TMEM106B S186 mice had no difference in level of TMEM106B protein relative to wild-type and that homozygous *Grn*$^{-/-}$ TMEM106B$^{S186/S186}$ mice did not have ameliorated lysosomal proliferation or microgliosis relative to *Grn*$^{-/-}$ mice, making it less likely that the rs3173615 variant is causal[14]. Moreover, rs1990622 was found to be more significant than rs3173615 following meta-analysis in a recent FTLD-TDP GWAS, which would be unexpected if the protective effect of the rs1990622 minor allele was due to its linkage with rs3173615.

We used publicly available expression and protein datasets to establish that the rs1990622 minor allele typically acts as a pQTL but not as an eQTL. This suggests that the protective effect at the *TMEM106B* locus is likely mediated by a genetic variant that acts after transcription to reduce TMEM106B protein levels. Because intronic and intergenic variants are not incorporated into the processed mRNA molecule it is less likely that such variants, including rs1990622, are mediating the protective effect of this allele. Furthermore, the pQTL finding suggests that the *TMEM106B* locus may exert its protective effect through an effect on protein availability rather than changes in function due to an amino acid change, reducing the likelihood that rs3173615 is the causal variant. That said, the Nicholson et al. study suggested that the TMEM106B S185 protein is less stable than wild-type in an *in vitro* setting and may result in reduced protein levels as a result[13].

Lastly, we found in our LRS dataset, as well as in ADSP, that the *TMEM106B* 3' UTR deletion is in higher LD with rs1990622 than rs3173615, which is consistent with a model in which the

respective significance of these two SNVs is related to their linkage with the *TMEM106B* 3' UTR deletion.

There are several possibilities for how the *TMEM106B* 3' UTR deletion could mediate a protective effect against FTLD-TDP pathogenesis. The pQTL evidence indicates that the minor allele at the *TMEM106B* locus decreases TMEM106B protein levels. The deletion may result in selective enrichment of an alternate transcript polyadenylation site, changing the 3' UTR. Such a change in the 3' UTR could disrupt protein binding, which may in turn decrease translational efficiency, alter the subcellular localization of the RNA, or impair protein routing to the endoplasmic reticulum[34]. Identifying a clear-cut mechanism linking the deletion to reduced TMEM106B protein levels (and increased progranulin protein levels) is still required to confirm that this is the causal variant at the locus.

In summary, we report a large, prevalent SV on *TMEM106B* that is in perfect LD with rs1990622 and near-perfect LD with rs3173615 in a large LRS dataset. LRS provides a valuable tool for detection of large genomic variants that can aid the interpretation of GWAS results and elucidate the genetic drivers of disease.

Conflicts of interest

The authors have no conflicting interests.

Data Availability

Sieberts eQTL meta-analysis: https://www.synapse.org/#!Synapse:syn17015233

Wingo eQTL meta-analysis: https://www.synapse.org/#!Synapse:syn31826294

GTEx eQTL summary statistics:

http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/imported/GTEx_V8/

CommonMind Consortium eQTL summary statistics:

https://www.synapse.org/#!Synapse:syn4622659

Banner pQTL summary statistics: https://www.synapse.org/#!Synapse:syn24847777

ARIC pQTL summary statistics: http://nilanjanchatterjeelab.org/pwas/

MetaBrain eQTL summary statistics: https://www.metabrain.nl

eQTLgen eQTL summary statistics: https://www.eqtlgen.org

DECODE pQTL summary statistics (Supplementary Tables):

https://www.nature.com/articles/s41588-021-00978-w#MOESM4

ROSMAP proteomics BA6 and BA37: https://www.synapse.org/#!Synapse:syn25335376

ROSMAP proteomics BA9: https://www.synapse.org/#!Synapse:syn25006657

Mount Sinai Brain Bank proteomics: https://www.synapse.org/#!Synapse:syn25006647

AMP-AD WGS & SNP array: https://dss.niagads.org/sample-sets/snd10011/

The NGS and LRS genomes will be made available in a research repository after publication.

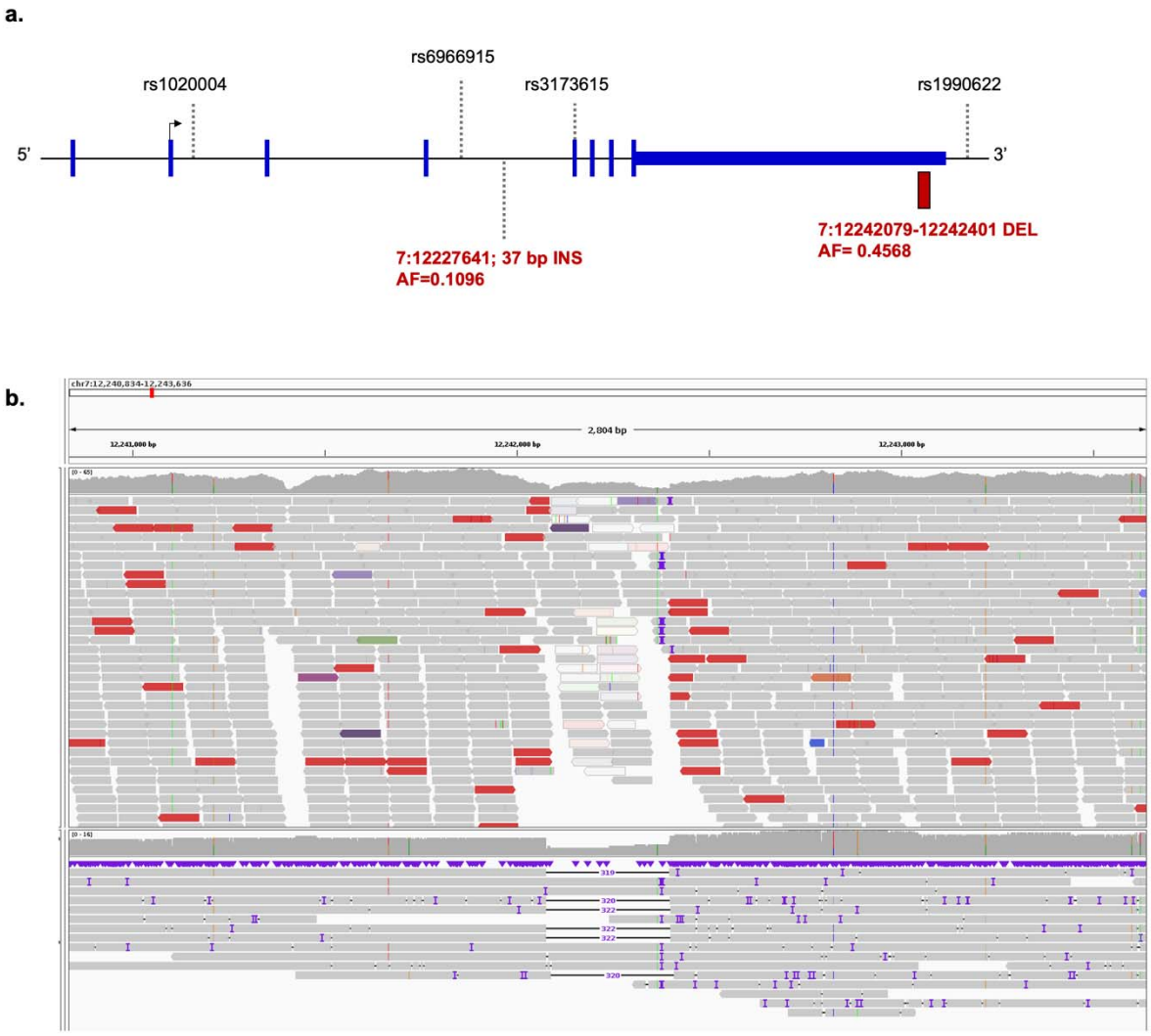

Figure 1. TMEM106B structural variants (a) Two structural variants were identified within 20 kilobases of *TMEM106B*. (b) The *TMEM106B* 3' UTR deletion was detected in both next-generation sequencing (top panel) and long-read sequencing (bottom panel). Both sequencing modalities are displayed here for a representative cohort participant carrying one copy of the *TMEM106B* 3' UTR deletion.

			eQTL		pQTL	
Dataset	**tissue**	**sample size**	**beta**	**p-value**	**beta**	**p-value**
GTEX	Blood	670	-	**0.0000**	-	-
GTEX	CTX	205	-	**0.0000**	-	-
Mayo	CER	275	0.3055	**0.0001**	-	-
Mayo	TCX	276	0.1529	0.0542	-	-
Wingo	DLPFC & PHG	722	0.0563	0.2416	-0.0773	0.1005
Metabrain	CTX	6601	0.0268	0.3115	-	-
eQTLGen	Blood	31427	-	0.0174	-	-
CommonMind	DLPFC	590	-0.0084	0.4044	-	-
Sieberts	CTX	1433	0.0260	0.4803	-	-
ROSMAP	DLPFC	269	-0.0214	0.7177	0.0093	**0.0117**
DECODE	Blood	35371	-	-	-0.1106	**0.0000**
Banner	DLPFC	129	-	-	-0.1786	**0.0000**
ROSMAP	DLPFC	116	-	-	0.0814	**0.0000**
ROSMAP	FC (BA6)	101	-	-	-0.1127	**0.0063**
ROSMAP	FC (BA9)	310	-	-	-0.2594	**0.0000**
MSBB	PHG	102	-	-	-0.3571	**0.0000**
ARIC AA	Plasma	1871	-	-	0.3147	**0.0000**
ARIC EA	Plasma	7213	-	-	-0.2547	**0.0000**
ROSMAP	TCX (BA37)	101	-	-	-0.1281	**0.0012**

Table 1. Effect of rs1990622 as a TMEM106B expression quantitative trait locus (eQTL) or protein quantitative trait locus (pQTL). Betas are those reported by respective studies and thus may be on different scales given different normalization procedures for transcriptomic and proteomic data.

Abbreviations: Cortex (CTX), cerebellum (CER), temporal cortex (TCX), DLPFC (dorsolateral prefrontal cortex), parahippocampal gyrus (PHG), frontal cortex (FC)

Acknowledgements

This work was supported by the following NIH grants: RO1AG060747; R35AG072290; R01AG048076; R01AG074339; P30AG066515; K99AG075238, and P30AG066515 Development project. This work was also supported by the Alzheimer's Association (AARF-20-683984).

Data for this study were prepared, archived, and distributed by the National Institute on Aging Alzheimer's Disease Data Storage Site (NIAGADS) at the University of Pennsylvania (U24-AG041689), funded by the National Institute on Aging

The Alzheimer's Disease Sequencing Project (ADSP) (NG00067) is comprised of two Alzheimer's Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. Seshadri and Fornage.

Sequencing for the Follow Up Study (FUS) is supported through U01AG057659 (to Drs. PericakVance, Mayeux, and Vardarajan) and U01AG062943 (to Drs. Pericak-Vance and Mayeux). Data generation and harmonization in the Follow-up Phase is supported by U54AG052427 (to Drs. Schellenberg and Wang). The FUS Phase analysis of sequence data is supported through U01AG058589 (to Drs. Destefano, Boerwinkle, De Jager, Fornage, Seshadri, and Wijsman), U01AG058654 (to Drs. Haines, Bush, Farrer, Martin, and Pericak-Vance), U01AG058635 (to Dr. Goate), RF1AG058066 (to Drs. Haines, Pericak-Vance, and Scott), RF1AG057519 (to Drs. Farrer and Jun), R01AG048927 (to Dr. Farrer), and RF1AG054074 (to Drs. Pericak-Vance and Beecham).

The ADGC cohorts include: Adult Changes in Thought (ACT) (U01 AG006781, U19 AG066567), the Alzheimer's Disease Research Centers (ADRC) (P30 AG062429, P30 AG066468, P30 AG062421, P30 AG066509, P30 AG066514, P30 AG066530, P30 AG066507, P30 AG066444, P30 AG066518, P30 AG066512, P30 AG066462, P30 AG072979, P30 AG072972, P30 AG072976, P30 AG072975, P30 AG072978, P30 AG072977, P30 AG066519, P30 AG062677, P30 AG079280, P30 AG062422, P30 AG066511, P30 AG072946, P30 AG062715, P30 AG072973, P30 AG066506, P30 AG066508, P30 AG066515, P30 AG072947, P30 AG072931, P30 AG066546, P20 AG068024, P20 AG068053, P20 AG068077, P20 AG068082, P30 AG072958, P30 AG072959), the Chicago Health and Aging Project (CHAP) (R01 AG11101, RC4 AG039085, K23 AG030944), Indiana Memory and Aging Study (IMAS) (R01 AG019771),

Indianapolis Ibadan (R01 AG009956, P30 AG010133), the Memory and Aging Project (MAP) (R01 AG17917), Mayo Clinic (MAYO) (R01 AG032990, U01 AG046139, R01 NS080820, RF1 AG051504, P50 AG016574), Mayo Parkinson's Disease controls (NS039764, NS071674, 5RC2HG005605), University of Miami (R01 AG027944, R01 AG028786, R01 AG019085, IIRG09133827, A2011048), the Multi-Institutional Research in Alzheimer's Genetic Epidemiology Study (MIRAGE) (R01 AG09029, R01 AG025259), the National Centralized Repository for Alzheimer's Disease and Related Dementias (NCRAD) (U24 AG021886), the National Institute on Aging Late Onset Alzheimer's Disease Family Study (NIA- LOAD) (U24 AG056270), the Religious Orders Study (ROS) (P30 AG10161, R01 AG15819), the Texas Alzheimer's Research and Care Consortium (TARCC) (funded by the Darrell K Royal Texas Alzheimer's Initiative), Vanderbilt University/Case Western Reserve University (VAN/CWRU) (R01 AG019757, R01 AG021547, R01 AG027944, R01 AG028786, P01 NS026630, and Alzheimer's Association), the Washington Heights-Inwood Columbia Aging Project (WHICAP) (RF1 AG054023), the University of Washington Families (VA Research Merit Grant, NIA: P50AG005136, R01AG041797, NINDS: R01NS069719), the Columbia University Hispanic Estudio Familiar de Influencia Genetica de Alzheimer (EFIGA) (RF1 AG015473), the University of Toronto (UT) (funded by Wellcome Trust, Medical Research Council, Canadian Institutes of Health Research), and Genetic Differences (GD) (R01 AG007584). The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193.

The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904), the EU Joint Programme – Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435. ARIC research is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was

provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS (R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their participants, faculty and staff. The content of these manuscripts is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services.

The FUS cohorts include: the Alzheimer's Disease Research Centers (ADRC) (P30 AG062429, P30 AG066468, P30 AG062421, P30 AG066509, P30 AG066514, P30 AG066530, P30 AG066507, P30 AG066444, P30 AG066518, P30 AG066512, P30 AG066462, P30 AG072979, P30 AG072972, P30 AG072976, P30 AG072975, P30 AG072978, P30 AG072977, P30 AG066519, P30 AG062677, P30 AG079280, P30 AG062422, P30 AG066511, P30 AG072946, P30 AG062715, P30 AG072973, P30 AG066506, P30 AG066508, P30 AG066515, P30 AG072947, P30 AG072931, P30 AG066546, P20 AG068024, P20 AG068053, P20 AG068077, P20 AG068082, P30 AG072958, P30 AG072959), Alzheimer's Disease Neuroimaging Initiative (ADNI) (U19AG024904), Amish Protective Variant Study (RF1AG058066), Cache County Study (R01AG11380, R01AG031272, R01AG21136, RF1AG054052), Case Western Reserve University Brain Bank (CWRUBB) (P50AG008012), Case Western Reserve University Rapid Decline (CWRURD) (RF1AG058267, NU38CK000480), CubanAmerican Alzheimer's Disease Initiative (CuAADI) (3U01AG052410), Estudio Familiar de Influencia Genetica en Alzheimer (EFIGA) (5R37AG015473, RF1AG015473, R56AG051876), Genetic and Environmental Risk Factors for Alzheimer Disease Among African Americans Study (GenerAAtions) (2R01AG09029, R01AG025259, 2R01AG048927), Gwangju Alzheimer and Related Dementias Study (GARD) (U01AG062602), Hillblom Aging Network (2014-A-004-NET, R01AG032289, R01AG048234), Hussman Institute for Human Genomics Brain Bank (HIHGBB) (R01AG027944, Alzheimer's Association "Identification of Rare Variants in Alzheimer Disease"),

Ibadan Study of Aging (IBADAN) (5R01AG009956), Longevity Genes Project (LGP) and LonGenity (R01AG042188, R01AG044829, R01AG046949, R01AG057909, R01AG061155, P30AG038072), Mexican Health and Aging Study (MHAS) (R01AG018016), Multi-Institutional Research in Alzheimer's Genetic Epidemiology (MIRAGE) (2R01AG09029, R01AG025259, 2R01AG048927), Northern Manhattan Study (NOMAS) (R01NS29993), Peru Alzheimer's Disease Initiative (PeADI) (RF1AG054074), Puerto Rican 1066 (PR1066) (Wellcome Trust (GR066133/GR080002), European Research Council (340755)), Puerto Rican Alzheimer Disease Initiative (PRADI) (RF1AG054074), Reasons for Geographic and Racial Differences in Stroke (REGARDS) (U01NS041588), Research in African American Alzheimer Disease Initiative (REAAADI) (U01AG052410), the Religious Orders Study (ROS) (P30 AG10161, P30 AG72975, R01 AG15819, R01 AG42210), the RUSH Memory and Aging Project (MAP) (R01 AG017917, R01 AG42210Stanford Extreme Phenotypes in AD (R01AG060747), University of Miami Brain Endowment Bank (MBB), University of Miami/Case Western/North Carolina A&T African American (UM/CASE/NCAT) (U01AG052410, R01AG028786), and Wisconsin Registry for Alzheimer's Prevention (WRAP) (R01AG027161 and R01AG054047).

The four LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), The American Genome Center at the Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University Genome Institute (U54HG003079). Genotyping and sequencing for the ADSP FUS is also conducted at John P. Hussman Institute for Human Genomics (HIHG) Center for Genome Technology (CGT).

Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Centralized Repository for Alzheimer's Disease and Related Dementias (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Study Investigators, the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC, U24AG072122) and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA. Harmonized phenotypes were provided by the ADSP Phenotype Harmonization Consortium (ADSP-PHC), funded by NIA (U24 AG074855, U01 AG068057 and R01 AG059716) and Ultrascale Machine Learning to Empower Discovery in Alzheimer's Disease Biobanks (AI4AD, U01 AG068057). This research was supported in part by the Intramural Research Program of the National Institutes of health, National Library of Medicine. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or nongovernmental organizations.

References

1. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016 Apr 15;32(8):1220–2.

2. Pedersen BS, Quinlan AR. Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. Gigascience [Internet]. 2019 Apr 1;8(4). Available from: http://dx.doi.org/10.1093/gigascience/giz040

3. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019 Nov 20;20(1):246.

4. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020 Oct;21(10):597–614.

5. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019 Apr 16;10(1):1784.

6. Trelle AN, Carr VA, Guerin SA, Thieu MK, Jayakumar M, Guo W, et al. Hippocampal and cortical mechanisms at retrieval explain variability in episodic remembering in older adults. Elife [Internet]. 2020 May 29;9. Available from: http://dx.doi.org/10.7554/eLife.55335

7. Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet. 2022 Apr;54(4):412–36.

8. Van Deerlin VM, Sleiman PMA, Martinez-Lage M, Chen-Plotkin A, Wang LS, Graff-Radford NR, et al. Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. Nat Genet. 2010 Mar;42(3):234–9.

9. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.

10. Finch N, Carrasquillo MM, Baker M, Rutherford NJ, Coppola G, Dejesus-Hernandez M, et al. TMEM106B regulates progranulin levels and the penetrance of FTLD in GRN mutation carriers. Neurology. 2011 Feb 1;76(5):467–74.

11. Cruchaga C, Graff C, Chiang HH, Wang J, Hinrichs AL, Spiegel N, et al. Association of TMEM106B gene polymorphism with age at onset in granulin mutation carriers and plasma granulin protein levels. Arch Neurol. 2011 May;68(5):581–6.

12. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015 Nov 1;31(21):3555–7.

13. Nicholson AM, Finch NA, Wojtas A, Baker MC, Perkerson RB 3rd, Castanedes-Casey M, et al. TMEM106B p.T185S regulates TMEM106B protein levels: implications for frontotemporal dementia. J Neurochem. 2013 Sep;126(6):781–91.

14. Cabron AS, Borgmeyer U, Richter J, Peisker H, Gutbrod K, Dörmann P, et al. Lack of a protective effect of the Tmem106b “protective SNP” in the Grn knockout mouse model for frontotemporal lobar degeneration. Acta Neuropathol Commun. 2023 Jan 27;11(1):21.

15. Pottier C, Zhou X, Perkerson RB 3rd, Baker M, Jenkins GD, Serie DJ, et al. Potential genetic modifiers of disease risk and age at onset in patients with frontotemporal lobar degeneration and GRN mutations: a genome-wide association study. Lancet Neurol. 2018 Jun;17(6):548–58.

16. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100.

17. Smolka M, Paulin LF, Grochowski CM, Mahmoud M, Behera S, Gandhi M, et al. Comprehensive Structural Variant Detection: From Mosaic to Population-Level [Internet]. bioRxiv. 2022 [cited 2023 Apr 17]. p. 2022.04.04.487055. Available from: https://www.biorxiv.org/content/10.1101/2022.04.04.487055v1.full

18. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples [Internet]. bioRxiv. 2018 [cited 2023 Jul 6]. p. 201178. Available from: https://www.biorxiv.org/content/10.1101/201178v3

19. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013 Mar;14(2):178–92.

20. English AC, McCarthy N, Flickenger R, Maheshwari S, Meed L, Mangubat A, et al. Leveraging a WGS compression and indexing format with dynamic graph references to call structural variants [Internet]. bioRxiv. 2020 [cited 2023 Apr 6]. p. 2020.04.24.060202. Available from: https://www.biorxiv.org/content/10.1101/2020.04.24.060202v1

21. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015 Feb 25;4:7.

22. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience [Internet]. 2021 Feb 16;10(2). Available from: http://dx.doi.org/10.1093/gigascience/giab008

23. Chen CY, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL. Improved ancestry inference using weights from external reference panels. Bioinformatics. 2013 Jun 1;29(11):1399–406.

24. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.

25. Sieberts SK, Perumal TM, Carrasquillo MM, Allen M, Reddy JS, Hoffman GE, et al. Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. Sci Data. 2020 Oct 12;7(1):340.

26. Hoffman GE, Bendl J, Voloudakis G, Montgomery KS, Sloofman L, Wang YC, et al. CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia

and Bipolar Disorder. Sci Data. 2019 Sep 24;6(1):180.

27. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020 Sep 11;369(6509):1318–30.

28. Wingo T, Liu Y, Gerasimov ES, Vattathil S, Liu J, Cutler D, et al. Integrating Human Brain Proteomes With GWAS Results to Identify Causal Brain Proteins for the Major Psychiatric Disorders. Biol Psychiatry. 2023 May 1;93(9, Supplement):S55–6.

29. de Klein N, Tsai EA, Vochteloo M, Baird D, Huang Y, Chen CY, et al. Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. Nat Genet. 2023 Mar;55(3):377–88.

30. Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021 Sep;53(9):1300–10.

31. Zhang J, Dutta D, Köttgen A, Tin A, Schlosser P, Grams ME, et al. Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. Nat Genet. 2022 May;54(5):593–602.

32. Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, et al. Long read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits [Internet]. bioRxiv. 2020 [cited 2023 Apr 14]. p. 848366. Available from: https://www.biorxiv.org/content/10.1101/848366v2

33. Robins C, Liu Y, Fan W, Duong DM, Meigs J, Harerimana NV, et al. Genetic control of the human brain proteome. Am J Hum Genet. 2021 Mar 4;108(3):400–10.

34. Mitschka S, Mayr C. Context-specific regulation and function of mRNA alternative polyadenylation. Nat Rev Mol Cell Biol. 2022 Dec;23(12):779–96.