Rare Variant Aggregation in 148,508 Exomes Identifies Genes Associated with Proxy Alzheimer's Disease

Douglas P Wightman[1], Jeanne E Savage[1], Christiaan A de Leeuw[1], Iris E Jansen[1], Danielle Posthuma[1]

Affiliations

1) Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU Amsterdam, Amsterdam, the Netherlands.

Abstract

We generated a proxy Alzheimer's disease phenotype for 148,508 individuals in the UK biobank in order to perform exome-wide rare variant aggregation analyses to identify genes associated with proxy Alzheimer's disease. We identified four genes significantly associated with the proxy phenotype, three of which have been previously associated with clinically diagnosed Alzheimer's disease (*SORL1*, *TREM2*, and *TOMM40*). We identified one gene (*HEXA*) which has not been previously associated with Alzheimer's disease but is known to contribute to neurodegenerative disease. Here we show that proxy Alzheimer's disease can capture some of the rare variant association signal for Alzheimer's disease and can be used to highlight genes and variants of interest. The proxy phenotype allows for the utilisation of large genetic databases without clinically diagnosed Alzheimer's disease patients to uncover variants and genes that contribute to Alzheimer's disease.

Introduction

Rare variants (minor allele frequency (MAF) <0.01) contributing to Alzheimer's disease (AD) have frequently been identified, first by family-based linkage studies[1–3], and later by exome sequencing[4,5] and whole gene sequencing[6,7]. Through these methods multiple genes have been reliably associated with AD through rare variants[8,9]. The samples sizes for these studies generally ranges from a few thousand individuals[7] to tens of thousands[10]. Studies with larger sample sizes are more likely to observe rarer variants which provides greater power to conduct rare-variant analyses. Ultra-rare (MAF<$1x10^{-4}$) variants are of particular interest because they are more likely to have a larger impact on the protein of interest[11], as deleterious variants are likely to undergo negative selection. Due to the relatively late-onset of AD, very few patients are included in large biobank cohorts so large clinically diagnosed AD cohorts have to be generated through patient recruitment. This process is time-consuming and financially costly. However, estimation of a proxy AD phenotype allows for the utilisation of large biobank cohorts to identify variants and genes associated with AD through rare variants. The first description of a proxy AD phenotype based on familial AD status was described in Liu *et al.* (2017)[12] and later a common variant driven genome-wide association meta-analysis was performed by Marioni *et al.* (2018)[13], which included a proxy AD phenotype for the UK biobank (UKB) participants. Both of these studies used a case-control design for the proxy phenotype.

Jansen *et al.* (2019)[14] generated a quantitative proxy phenotype for UKB participants in order to include them with clinically defined cases and controls in a genome-wide meta-analysis of common variation in AD. The proxy phenotype was calculated by assigning a quantitative phenotype to individuals based on their number of parents diagnosed with AD. This proxy phenotype is a more dilute phenotype compared to a phenotype based on clinical diagnosis because genetic risk variants of AD can be lost when alleles are transmitted from parent to offspring. Jansen and colleagues[14] showed that the power lost due to the diluted phenotype was compensated for by the large sample size of 376,113 individuals. The genetic correlation between the proxy AD and clinically diagnosed AD was high (r_g = 0.81) so the proxy phenotype was able to capture a good amount of the genetic signal of AD. In the current study, we aimed to apply this same proxy phenotype to rare variant analyses to determine if the proxy phenotype can recapture AD associated variants and genes using rare-variants. In the process, we aim to identify additional genes and variants of interest to AD.

Results

Variant aggregation analysis

We performed genome-wide gene-level variant aggregation analyses for 4 variant categories in 148,508 individuals of European ancestry with a proxy score for AD. Of the 148,508 individuals, 22,080 individuals had at least one parent with AD or an AD diagnosis themselves. Variant aggregation analyses using different variant categories were chosen to limit the analyses to variants with high predicted impacts on protein function to aid interpretation of significant associations. The four categories were based on the predicted impact of the variant on the protein, with the most impactful variants identified as high confidence predicted loss-of-function (HiC pLOF) and the least impactful variants identified as missense. The two intermediate categories were predicted loss-of-function (pLOF) and high confidence missense variants defined by a REVEL score >50 (REVEL>50). The four tested categories were HiC pLOF variants alone (HiC pLOF), all pLOF variants (pLOF), all pLOF variants plus high confidence missense variants (pLOF+REVEL>50), and all pLOF variants plus all missense variants (pLOF+missense).

The number of variants in each variant category grows as less impactful variants are included, this also increases the number of tested genes where at least one variant maps to the gene (**Table 1**). The power to observe associations increased with increased number of variants and this was reflected in the increasing genomic inflation factors as each variant category increases the number of tested genes and variants **(Supplementary Table 1)**. The genomic inflation factors of the variant aggregation analyses were less than one in all four variant categories (HiC pLOF=0.9079775, pLOF=0.9224255, pLOF+REVEL>50=0.9278713, pLOF+missense=0.9443691). The low genomic inflation factors suggest that the signal within the data is sparse, especially in the pLOF variant categories where numbers of variants within a gene are low. However, four unique genes were found to be significantly associated with proxy AD (**Figure 1**).

Significant Genes

TREM2, *TOMM40*, *SORL1*, and *HEXA* were the four unique genes which reached significance in the SKAT-O[15] variant aggregation analyses after Bonferroni correction for the number of genes and variant categories **(Table 2).** *TREM2*, *TOMM40*, and *SORL1* have been identified in previous rare variant and common variant genome-wide association studies of AD[8,14,16]. *HEXA* has not been previously identified in any rare variant or common variant genome-wide association study of AD. No significantly associated genes were identified with either HiC pLOF or all pLOF variants alone. *SORL1* and *HEXA* were significantly associated genes when testing all pLOF and high confidence missense variants (pLOF+REVEL >50). *TREM2* and *TOMM40* were significantly associated when testing all pLOF and missense variants (pLOF+missense).

TREM2 was significantly associated with the phenotype in the pLOF+missense variant category (P=2.45x10^{-11}, Nsnps=117). The significance of this gene was largely due to two rare missense variants (rs75932628: MAF=0.003081, P=2.55x10^{-9}; rs143332484: MAF=0.009787, P=4.37x10^{-5}) and one ultra-rare high-confidence missense variant (6:41161292:A_C: MAF=3.37x10^{-6} , P=7.47x10^{-6}) **(Figure 2; Supplementary Table 2)**. The most significant variant (rs75932628) is a genome-wide significant missense variant that causes an amino acid substitution from arginine to histidine (p.R239W) in exon 2 of *TREM2*.

This variant has a CADD score of 26.1 and was identified as the most associated variant in *TREM2* in Sims *et al*. (2017)[17]. The other missense variant (rs143332484) also causes an amino acid substitution from arginine to histidine (p.R47H) in exon 2 of *TREM2*, however this variant is predicted to be less deleterious (CADD=9.7). rs143332484 was also found to be significant in Sims *et al*. (2017)[17]. Both of these variants have been identified to cause reduced ligand affinity, signalling response, and phagocytosis of lipoprotein in microglia[18]. The high-confidence missense variant (6:41161292:A_C) also impacts exon 2 of *TREM2* and causes an amino acid substitution from leucine to arginine (p.L121R). This substitution affects a conserved region and has a CADD score of 28.3. *TREM2* is a well characterised AD gene and encodes a protein which is known to impact microglia function and inflammation[18].

TOMM40 was significantly associated with proxy AD when all missense and pLOF variants were included in the variant aggregation analysis (P=2.39x10^{-10}, Nsnps=119). The gene association is largely driven by one significant rare missense variant (rs142412517; MAF=0.001731; P=4.43x10^{-10}) **(Figure 2; Supplementary Table 2)**. This missense variant (CADD=25.4) causes an amino acid change from arginine to tryptophan (p.R239W) in a conserved region but the impact on the protein is unclear. This lead variant is in very low LD (<0.01) with the APOE e4 variants (rs7412-C, rs429358-C), but when conditioning for APOE e4 status based on these two variants *TOMM40* was no longer significantly associated with the phenotype (P=4.28x10^{-2}) **(Supplementary Table 3)**. No other significant gene lost significance when conditioned on APOE e4 status. Of the 119 variants which mapped to *TOMM40,* 106 were missense variant, 10 were high-confidence missense variants, and 3 were high-confidence pLOF variants. The low number of variants in the HiC LOF, pLOF, and pLOF+REVEL>50 analyses caused this gene to fail the cumulative frequency threshold (0.0001) so *TOMM40* was not included in these analyses. The influence of missense variants and *APOE* e4 variants on the association of *TOMM40* suggests that the impact of the variant on *TOMM40* protein structure is not the primary driver of the association.

SORL1 was significantly associated with proxy AD when restricting to pLOF and high confidence missense variants (P=1.75x10^{-9}, Nsnps=279). *SORL1* is a known AD gene and encodes a protein important in amyloid precursor protein processing[11]. The association signal was mostly driven by 1 ultra-rare pLOF variant (11:121520795:A_G: MAF= 3.37x10^{-6}, P= 1.12x10^{-5}), two ultra-rare high confidence missense variants (rs780053569: MAF=6.73x10^{-6}, P=4.99x10^{-6}; rs139351633: MAF= 2.69x10^{-5}, P= 4.75x10^{-5}), and one rare high confidence missense variant (rs140327834: MAF= 0.003768, P= 9.81x10^{-5}) (**Figure 3; Supplementary Table 2**). The most significant variant (rs780053569) is a missense variant (CADD=26.2) which causes an amino acid substitution from serine to leucine (p.S602L) in a conserved region. The second most significant variant (rs371418037) is a missense variant (CADD=24.2) which causes an amino acid substitution from glutamine to histidine (p.Q614H) in a conserved region and is predicted to affect protein stability. The third most significant variant (11:121520795:A_G) is a pLOF variant (CADD=39) which causes the premature introduction of a stop codon (p.W450*) in exon 9. The other variants also cause amino acid substitutions but are not predicted to impact the protein function. The top 3 most significant variants in (rs780053569, rs371418037, 11:121520795:A_G) cause amino acid substitutions in the amyloid-beta binding region (vacuolar protein sorting 10 domain) of *SORL1*[11]. rs139351633 and rs140327834 affect the fibronectin type III and LDL-receptor class

B regions respectively. Holstege *et al*. (2017)[11] outlines the predictive criteria for *SORL1* variants and suggests that extremely rare variants (MAF <$1x10^{-4}$) with CADD scores >30 are likely pathogenic variants and truncating variants are pathogenic variants. 11:121520795:A_G fulfils all of these criteria suggesting that it is a variant of particular interest in *SORL1*. rs140327834 is the only highlighted *SORL1* variant mentioned in Holstege *et al*. (2017)[11], and was categorised as likely benign. The absence of the other highlighted variants is likely due to their frequency being too low to be identified in the smaller sample (~5000 individuals).

HEXA was significantly associated with the phenotype when restricting to pLOF and high confidence missense variants (*P*= $2.13x10^{-7}$, Nsnps=121). The association was mostly driven by one ultra-rare pLOF variant (15:72344078:C_G: MAF=$3.37x10^{-6}$, *P*=$1.16x10^{-5}$), one rare high-confidence missense variant (rs121907970: MAF=0.003879, *P*=$2.07x10^{-6}$), and one ultra-rare high-confidence missense variant (rs2088635643: MAF= $3.37x10^{-6}$, *P*=$1.26x10^{-5}$) (**Figure 2; Supplementary Table 2**). The most significant variant (rs121907970) is a missense variant (CADD=25.1) which causes an amino acid substitution from arginine to tryptophan (R247W/R258W). This amino acid change has been linked to a reduced activity of Hex A, however this reduction in activity is only in regards to an artificial substrate used to measure Hex A activity and does not affect in situ activity[19]. rs2088635643 is also a missense variant (CADD=31) which causes and amino acid substitution for tryptophan to arginine (W373R/W384R). However, neither of these variants are predicted to impact the protein function. The pLOF variant (15:72344078:C_G) causes a stop loss (*530S/*541S) in the last exon of *HEXA* (exon 14) but the consequence of this mutation is unknown. *HEXA*, along with *HEXB*, encodes the protein Hex which mediates breakdown of GM2 gangliosides, the accumulation of which contributes to "progressive neurodegenerative diseases, such as Sandhoff disease (SD) and Tay-Sachs disease (TSD)"[20]. There is also some evidence to support the colocalization of Hex with amyloid-beta plaques in the brains of AD patients[21]. At least one risk allele of the 3 highlighted *HEXA* variants (15:72344078:C_G, rs2088635643, rs121907970) was present in 6 of the 277 (2.17%) individuals with an AD diagnosis, 16 of the 1075 (1.49%) individuals with two parents with an AD diagnosis, 195 of the 20,728 (0.94%) individuals with one parent with an AD diagnosis, and 933 of the 126,428 (0.74%) individuals with no AD diagnosis and no parent with an AD diagnosis.

Quality metrics (mean sequencing depth and variance, missing rate, and allele quality) of the variants highlighted above are available in Supplementary Table 2. Among the highlighted variants the mean read depth ranged from ~17-27, the missing rate was less than $1x10^{-4}$, and the allele quality (phred scale) ranged from 36-58 (approximately 99.9-99.999% base call accuracy). Quality metrics for all variants which were included in the analysis where *HEXA* was significantly associated with the proxy phenotype are available in Supplementary Table 4. Among the *HEXA* variants (N=121) the mean depth ranged from ~15-36, the missing rate was less than $1x10^{-3}$, and the allele quality (phred scale) ranged from 36-58.

Replication of known AD loci

We successfully replicated the rare variant based association of *TREM2*, *SORL1*, and the *APOE* locus (*TOMM40*). However, we were unable to replicate previous associations with *APP*, *PSEN1*, *PSEN2*, *ABCA7*, *BIN1*, *UNC5C*, *AKAP9*, *NOTCH3*, *CLU*, *PLGC2*, and *ABI3*

(**Supplementary Table 5**)[8,9]. Only three of these genes have nominal significance in any of the variant aggregation analyses (pLOF+missense: *ABCA7*: P=1.56x10^{-3}; *ABI3*: P=4.3x10^{-3}; *PSEN1*: P=0.04). *ABCA7*, *SORL1*, and *TREM2* are the most frequently reported genes in rare-variant association studies of AD[8], especially in recent exome-wide associations of unrelated individuals[4,5,10,22]. We identified two of the three most frequently identified genes in recent exome-wide association studies in unrelated individuals (*TREM2* and *SORL1*) and the third gene (*ABCA7*) was nominally significant.

Discussion

We performed a series of genome-wide variant aggregation analyses in 148,508 individuals with a proxy score for AD to successfully identify 3 known AD genes (*TREM2*, *TOMM40*, and *SORL1*) and one gene not previously associated with AD (*HEXA*). The analyses highlighted variants of interest in *TREM2*, *TOMM40*, *SORL1* and *HEXA*, where all of the highlighted variants caused an impact on the amino acid sequence. The minor alleles of all of the highlighted variants were positively associated with an increased proxy AD score. The role of *TREM2* and *SORL1* is well known in AD[9,11,18] and the loss of significance of *TOMM40* after condition on APOE e4 alleles suggests that the association of *TOMM40* is connected to the well-established *APOE* locus. *HEXA* is a known risk gene for Tay-Sachs disease[23], another neurodegenerative disease, and the protein it encodes has been colocalized with amyloid-beta plaques in the brains of AD patients[21]. Individuals with an AD diagnosis themselves had the highest frequency (2.17%) of at least one risk allele of the highlighted *HEXA* variants, which suggests that true AD cases are contributing to the *HEXA* association. This evidence suggests that *HEXA* may be an AD risk gene, however further analysis in a larger clinically diagnosed group is needed to confirm this association.

Across these analyses, we showed that the proxy phenotype can capture some AD gene associations identified from rare variant gene association studies of clinically diagnosed AD patients (*SORL1* and *TREM2*)[4,5,10,22]. However, we failed to identify further genes identified in previous rare variant analyses in clinically diagnosed AD patients[8]. This may be due to the differing study design between this current study and the design of the studies which initially identified these genes. The previously identified genes were found across different studies with different designs, including studies of family cohorts, unrelated individuals, exome-wide sequencing, whole gene sequencing, individual variant discovery, and variant aggregation. Previous exome-wide association studies of AD which performed similar variant aggregation as this study identified *ABCA7*, *SORL1*, and *TREM2*[4,10] as genes associated with AD. We were able to replicate the association of *SORL1* and *TREM2*, but not *ABCA7*.

A limitation of the proxy phenotype is that it is a less well-defined phenotype largely based on self-reported data. The question used to define parental status does not distinguish between AD and dementia, which introduces heterogeneity in the phenotype definition. The interpretation of proxy AD associated genes not previously associated with clinically diagnosed AD should be cautious as the heterogeneity of the phenotype may highlight dementia related genes rather than AD specific genes. However, a strength of the proxy phenotype is that it allows for the inclusion of more individuals in the study by utilising biobanks. The inclusion of more individuals is particularly beneficial to rare variant analyses as it increases the likelihood of identifying rare variants with large impacts on protein function. The results of this study support the conclusion from previous studies[12–14,24] which found that the proxy phenotype can capture AD genetic contributors and provide value to genetic studies of AD. We extend this to show that the proxy phenotypes can capture rare variants of interest to AD. It is important to note that variants and genes associated with proxy AD should be further validated in clinically diagnosed AD cohorts and proxy phenotypes are complementary to, and not a substitute for, well powered studies in clinically diagnosed cohorts.

Methods

Sample overview

This study performed exome-wide variant aggregation analyses using genetic data from 148,508 UKB participants of European ancestry. The UKB is a large population-based biobank which includes 503,325 individuals[25]. Individuals were selected for participation between 2006 and 2010. Invited individuals were between 40 and 69 years old, registered with the National Health Service, and living within 25 miles of one of the study research centres. Various data were collected from the individuals, including questionnaire answers, medical records, and genetic data. Of the 148,508 participants included in this analysis, 81,835 were female (55.1%), 66,673 were male (44.9%) and the median age was 58. All participants provided written informed consent; the UKB received ethical approval from the National Research Ethics Service Committee North West-Haydock (reference 11/NW/0382), and all study procedures were in accordance with the World Medical Association for medical research. Access to the UK Biobank data was obtained under application number 16406.

Phenotype Definition

The proxy phenotype is a quantitative phenotype ranging from 0-2, where individuals with a higher score are considered to be at higher risk of developing AD based on their own diagnosis and the diagnoses of their parents. The construction of the proxy score has been described previously[14]. In brief, individuals in the UKB that report an “Alzheimer's disease/dementia” diagnosis in either parent (data fields 20107 and 20110) are given a phenotype value based on the number of parents who have had a diagnosis. Individuals who report an AD diagnosis for themselves or have medical records reporting an AD diagnosis (ICD10 codes G30, G300, G301, G308, G309, F00, F000, F001, F002, F009 in data fields 41270, 41202 and 41204) are given the same score as individuals with two parents with AD diagnoses. The contributions of parents without AD diagnoses were weighted by their age (100-age of parent/100; capped at 0.32 per parent), with older parents without AD down-weighted relative to younger ones. This resulted in 126,428 individuals with no AD diagnoses and no affected parents (proxy score <1), 20,728 individuals with one affected parent, 1075 individuals with two affected parents, and 277 individuals with an AD diagnosis themselves.

Variant sequencing and quality control

Exome sequencing was performed for 200,643 participants of the UK Biobank study by a partnership of eight biopharmaceutical companies[26,27]. Sequencing occurred in two batches, with the first ~50,000 individuals (UKB 50k) selected for completeness of phenotypic data and the presence of respiratory disorders of interest, and the next batch (UKB 150k) randomly selected from the larger sample of ~500,000 individuals. Targeted regions of the exome (39Mbp in total, including 100bp flanking each gene target) were captured using the IDT xGen Exome Research Panel v1.0 with dual-indexed 75 x 75 bp paired-end reads on the Illumina NovaSeq 6000 platform using S2 (UKB 50k) and S4 (UKB 150k) flow cells.

Raw sequence reads were mapped to the GRCh38 reference genome using the OQFE protocol, followed by duplicate read marking, variant calling with DeepVariant, and

filtering/merging with GLnexus. Full details of the protocol and settings are provided by Szustakowski *et al*. (2021)[26]. The resulting joint variant call file released by UKB included a total of 17,981,897 variants, with greater than 20x average coverage of 95.6% of sites in the target region. Our inspection of the data showed consistency with quality control recommendations[28] (e.g. all samples had a transition/transversion [Ti/Tv] ratio between 2.96 and 3.21 [*M* = 3.05] for known variants; all samples had between 47,000 and 72,000 total SNPs [*M* = 54,932] in the targeted plus flanking regions). We additionally filtered the set of variants released by UKB to exclude autosomal variants with missingness > 5% (n = 343,110), variants with a minor allele count (MAC) of 0 (n = 81,027), duplicates (n = 28,005), and variants outside of the targeted exome capture regions for which coverage and other quality metrics were not optimized (n = 8,800,694). After quality control, a total of 8,700,920 variants were available, of which 6,805,307 rare variants (MAF < .01) were selected for annotation.

Array-based genotypes were also available from these same samples[29]. We used indicators of genetic kinship from these data, as provided by UKB (field id 22021) to exclude 3rd degree or closer relatives. We also used these genotypes to empirically assign individuals to ancestral continental populations based on their similarity to the 1000 Genomes reference panel ancestries, and to calculate within-ancestry principal components, as described in detail by Jansen *et al.* (2019)[14]. Sample exclusion based on relatedness, ancestry, and withdrawn subjects resulted in 159,660 participants of European ancestry with exome sequencing data available for analysis. A proxy phenotype could be calculated for 148,508 individuals with available exome sequence data and these individuals were used in the analyses.

Variant annotation with VEP

Variants were annotated with ensembl variant effect predictor (VEP) v100.4[30] using ensembl version 100 data. pLOF variants were annotated using the LOFTEE plugin[31] (github commit 2df8880). Missense variants were annotated using the REVEL v1.3 plugin[32]. Variant categories were determined based on the predicted impact of the minor allele on the gene. Four categories were created; HiC pLOF, pLOF, pLOF+REVEL>50, and pLOF+missense. HiC pLOF represents the variants deemed as high confidence loss-of-function variants by LOFTEE. pLOF represents all predicted loss-of-function variants identified by any of the following gene consequences: start_lost, stop_lost, frameshift_variant, stop_gained, splice_donor_variant, splice_acceptor_variant, or transcript_ablation. pLOF+REVEL includes all of the variants in the pLOF category plus missense variants with a REVEL score >50. This REVEL threshold was chosen because it captured 75% of disease variants and ~11% of neutral variants in Ioannidis *et al.* (2016)[32]. The pLOF+missense category included all of the pLOF variants and all missense variants with a REVEL >=0.

Variant aggregation analyses

Variants were aggregated in a SKAT-O analysis, an optimal unified test which combines burden and kernel-based tests to maximise power[15]. Only rare variants (MAF<0.01) were included in the analyses. The variants were aggregated with default weights within their mapped genes. Four SKAT-O analyses were performed, one for each of the variant categories (HiC pLOF, pLOF, pLOF+REVEL>50, and pLOF+missense). These four categories were chosen to restrict the variant aggregation to variants with likely impact on

genes to aid interpretation of significantly associated genes. The analyses were performed using batch (UKB50k vs UKB150k), sex, age, and the first 10 ancestry principal components as covariates. Genes with low cumulative allele frequency (<0.0001) were removed to prevent genes with very few variants from biasing the analyses. The genomic inflation factors were calculated based on the P-values of each analysis. Genes were considered significant after Bonferroni correction for the number of genes in each test and the number of times the gene was tested across the four categories.

Investigating single variants

In order to highlight variants of interest, PLINK (v1.9b_6.17)[33,34] linear regression was performed to identify the association of the variants within significant genes identified in the SKAT-O analyses (*TREM2*, *TOMM40*, *SORL1* and *HEXA*) with the proxy phenotype. The variant association analysis was limited to the variants included in the SKAT-O analyses. The linear regression was performed using batch, sex, age, and the first 10 principal components as covariates. Variants with moderate associations ($P<1x10^{-4}$) were chosen as variants of interest. The impact of the variants on the proteins was predicted using mutfunc[35]. CADD[36] score and amino acid substitutions were identified during the initial variant annotation by VEP. All plots were generated in R using ggplot2[37] or base R[38] functions.

References

1. Cruts, M. *et al.* Estimation of the Genetic Contribution of Presenilin-1 and -2 Mutations in a Population-Based Study of Presenile Alzheimer Disease. *Hum. Mol. Genet.* **7**, 43–51 (1998).
2. Goate, A. *et al.* Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **349**, 704–706 (1991).
3. Pericak-Vance, M. A. *et al.* Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *Am. J. Hum. Genet.* **48**, 1034–1050 (1991).
4. Bis, J. C. *et al.* Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. *Mol. Psychiatry* **25**, 1859–1875 (2020).
5. Fan, K.-H. *et al.* Whole-Exome Sequencing Analysis of Alzheimer's Disease in Non-APOE*4 Carriers. *J. Alzheimers. Dis.* **76**, 1553–1565 (2020).
6. Vardarajan, B. N. *et al.* Rare coding mutations identified by sequencing of Alzheimer disease genome-wide association studies loci. *Ann. Neurol.* **78**, 487–498 (2015).
7. Jonsson, T. *et al.* Variant of TREM2 associated with the risk of Alzheimer's disease. *N. Engl. J. Med.* **368**, 107–116 (2013).
8. Hoogmartens, J., Cacace, R. & Van Broeckhoven, C. Insight into the genetic etiology of Alzheimer's disease: A comprehensive review of the role of rare variants. *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.* **13**, e12155 (2021).
9. Lord, J., Lu, A. J. & Cruchaga, C. Identification of rare variants in Alzheimer's disease. *Front. Genet.* **5**, 369 (2014).
10. Holstege, H. *et al.* Exome sequencing identifies rare damaging variants in the ATP8B4 and ABCA1 genes as novel risk factors for Alzheimer's Disease. *medRxiv* 2020.07.22.20159251 (2021) doi:10.1101/2020.07.22.20159251.
11. Holstege, H. *et al.* Characterization of pathogenic SORL1 genetic variants for association with Alzheimer's disease: a clinical interpretation strategy. *Eur. J. Hum. Genet.* **25**, 973–981 (2017).
12. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case-control association mapping by proxy using family history of disease. *Nat. Genet.* **49**, 325–331 (2017).
13. Marioni, R. E. *et al.* GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 99 (2018).
14. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
15. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
16. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
17. Sims, R. *et al.* Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nat. Genet.* **49**, 1373–1384 (2017).
18. Ulrich, J. D., Ulland, T. K., Colonna, M. & Holtzman, D. M. Elucidating the Role of TREM2 in Alzheimer's Disease. *Neuron* **94**, 237–248 (2017).
19. Mistri, M. *et al.* Identification of Novel Mutations in HEXA Gene in Children Affected with Tay Sachs Disease from India. *PLoS One* **7**, e39122 (2012).
20. Jia, M. *et al.* Microglia-Specific Expression of HEXA and HEXB Leads to Poor Prognosis

in Glioblastoma Patients. *Front. Oncol.* **11**, 685893 (2021).
21. Cataldo, A. M., Barnett, J. L., Mann, D. M. A. & Nixon, R. A. Colocalization of Lysosomal Hydrolase and β-Amyloid in Diffuse Plaques of the Cerebellum and Striatum in Alzheimer's Disease and Down's Syndrome. *J. Neuropathol. Exp. Neurol.* **55**, 704–715 (1996).
22. Raghavan, N. S. *et al.* Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer's disease. *Ann. Clin. Transl. Neurol.* **5**, 832–842 (2018).
23. Dersh, D., Iwamoto, Y. & Argon, Y. Tay-Sachs disease mutations in HEXA target the α chain of hexosaminidase A to endoplasmic reticulum-associated degradation. *Mol. Biol. Cell* **27**, 3813–3827 (2016).
24. Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).
25. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).
26. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.* **53**, 942–948 (2021).
27. Van Hout, C. V *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
28. GATK TEAM. Evaluating the quality of a germline short variant callset. https://gatk.broadinstitute.org/hc/en-us/articles/360035531572. Accessed 2021-10-14.
29. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
30. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
31. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
32. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
33. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
34. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, (2015).
35. Wagih, O. *et al.* A resource of variant effect predictions of single nucleotide variants in model organisms. *Mol. Syst. Biol.* **14**, e8430 (2018).
36. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2018).
37. Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).
38. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* (2017).

Acknowledgements

DP was funded by The Netherlands Organization for Scientific Research (NWO VICI 453-14-005), NWO Gravitation: BRAINSCAPES: A Roadmap from Neurogenetics to Neurobiology (Grant No. 024.004.012), and a European Research Council advanced grant (Grant No, ERC-2018-AdG GWAS2FUNC 834057). DW was funded by NWO Gravitation: BRAINSCAPES: A Roadmap from Neurogenetics to Neurobiology (Grant No. 024.004.012). IEJ was funded by NWO Gravitation: BRAINSCAPES: A Roadmap from Neurogenetics to Neurobiology (Grant No. 024.004.012). JES was supported by funding from the Amsterdam Neuroscience Alliance Project. CdL was funded by F. Hoffmann-La Roche AG. The research has been conducted using the UK Biobank Resource (application no. 16406). Analyses were carried out on the Genetic Cluster Computer hosted by the Dutch National computing and Networking Services SurfSARA.

Figures

Figure 1: Manhattan plot of the four variant aggregation analyses highlights four significantly associated genes (*TREM2*, *SORL1*, *HEXA*, and *TOMM40*). Each point represents a gene in one of the four variant aggregation analyses (HiCpLOF, pLOF, pLOF+REVEL>50, pLOF+missense). The dashed line represents the threshold of significance after correction for all genes tested across all four variant categories.

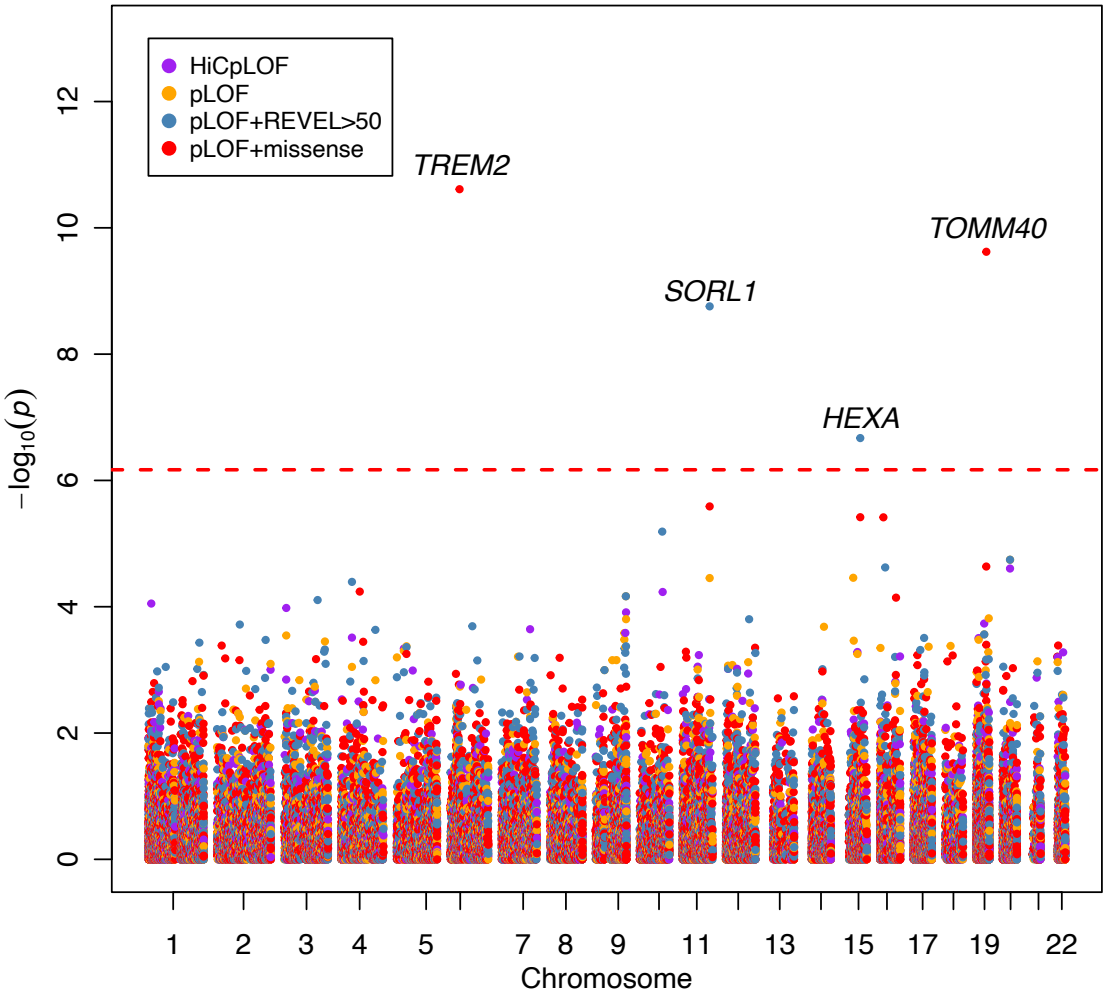


Figure 2: The single variant associations of variants which mapped to *TREM2* and TOMM40. Each point represents a variant and is coloured based on functional annotation (HiCpLOF, pLOF, REVEL>50, missense).

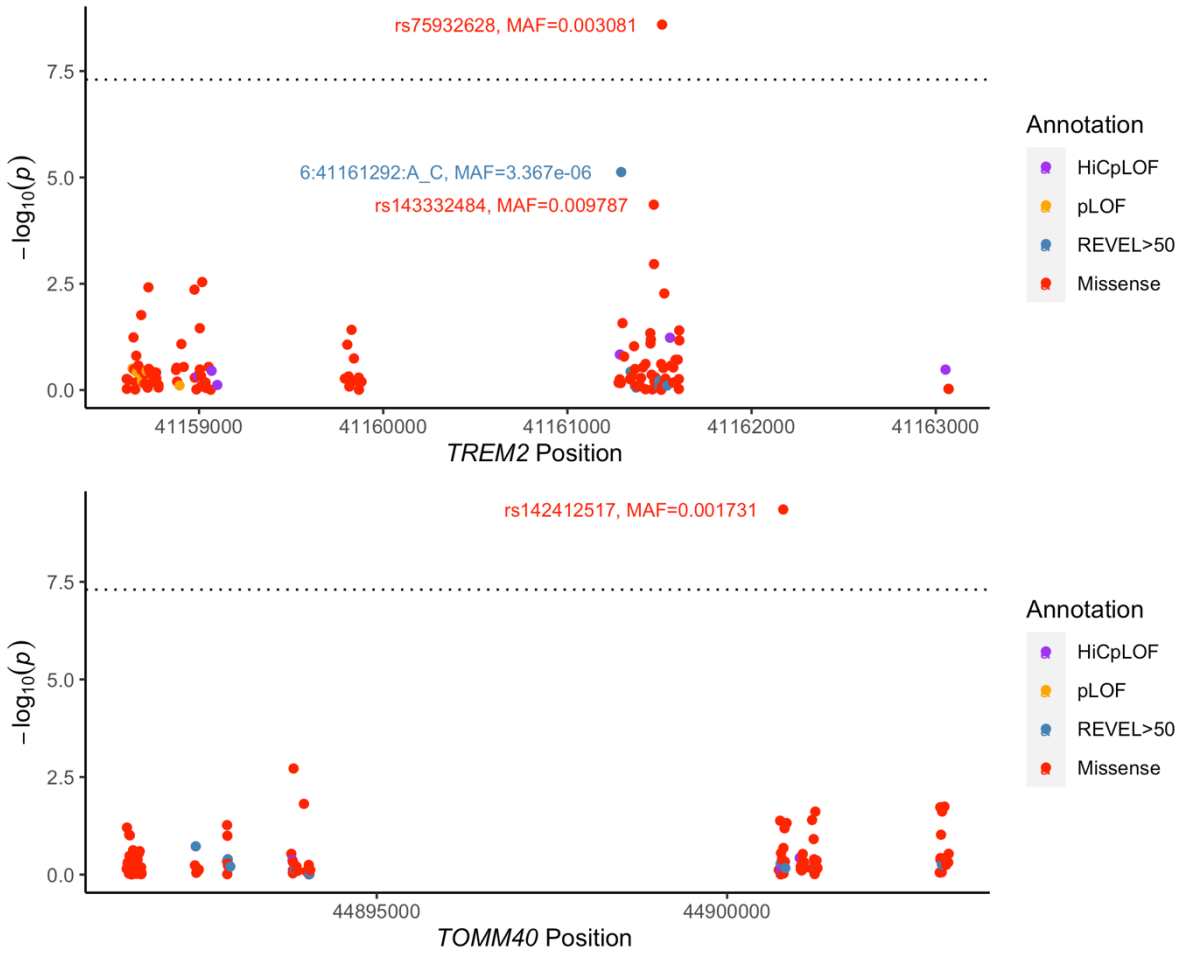
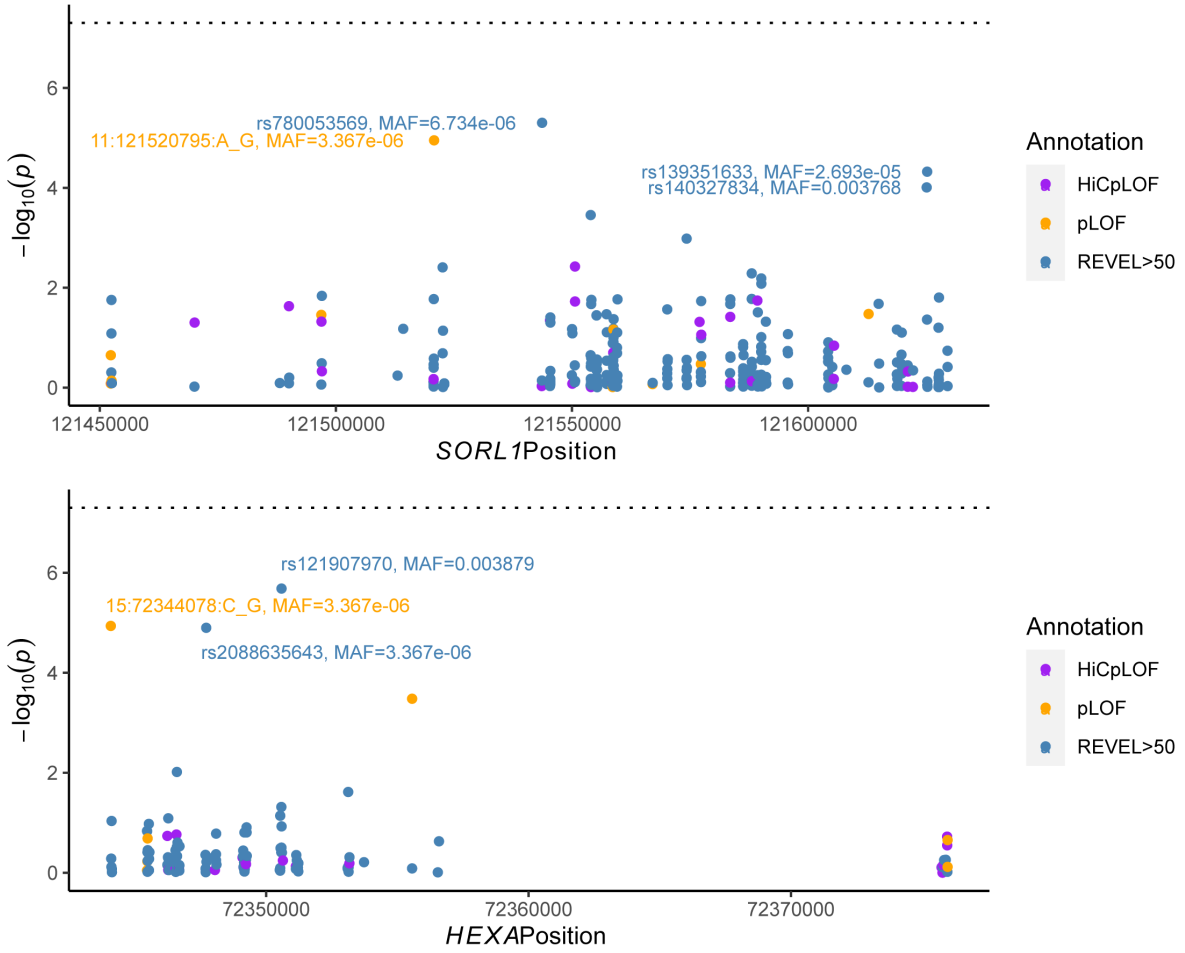


Figure 3: The single variant associations of variants which mapped to *SORL1* and *HEXA*. Each point represents a variant and is coloured based on functional annotation (HiCpLOF, pLOF, REVEL>50).



Tables

Table 1: The number of variants and genes included in each variant aggregation analysis. The numbers in brackets represent the potential number of genes in the analysis if genes were not excluded due to cumulative allele frequency filtering (<0.0001).

	HiC pLOF	pLOF	pLOF + REVEL >50	pLOF+ missense
# GENES	6749 (15572)	11038 (18071)	16010 (18508)	18425 (18605)
# VARIANTS	179501	295073	869322	3712943

Table 2: The results from the variant aggregation analyses (SKAT-O) of the four genes (*SORL1*, *HEXA*, *TREM2*, and *TOMM40*) which reached significance across the four variant categories (HiC pLOF, pLOF. pLOF+REVEL>50, and pLOF+missense).

	HiC pLOF		pLOF		pLOF+REVEL>50		pLOF+missense	
Gene	*P*	#variants	*P*	#variants	*P*	#variants	*P*	#variants
SORL1	0.0029	29	3.52×10^{-5}	39	1.75×10^{-9}**	279	2.58×10^{-6}*	782
HEXA	0.53	18	0.18	25	2.13×10^{-7}**	121	3.82×10^{-6}	196
TREM2	NA	6	0.47	12	0.22	19	2.45×10^{-11}**	117
TOMM40	NA	3	NA	3	NA	13	2.39×10^{-10}**	119

Significant after Bonferroni correction for number of genes within analysis*
Significant after Bonferroni correction for number of genes across all analyses **