# **Chapter 3 Supplementary Note**

**Rare Variant Aggregation in 148,508 Exomes Identifies Genes Associated with Proxy Dementia**

Table of Contents

[Chapter 3 Supplementary Note 1](#_Toc138599731)

[Supplementary Results 3](#_Toc138599732)

[Subthreshold genes 3](#_Toc138599733)

[Gene-set analyses 4](#_Toc138599734)

[*TOMM40* variants 5](#_Toc138599735)

[*HEXA* variants 5](#_Toc138599736)

[Comparison to previously identified AD genes through common variants 6](#_Toc138599737)

[Association of known AD genes in true case-control analysis 6](#_Toc138599738)

[*ABCA7* variants 7](#_Toc138599739)

[Power analysis 7](#_Toc138599740)

[Supplementary Methods 8](#_Toc138599741)

[Subthreshold genes 8](#_Toc138599742)

[SKAT-O gene-set analyses 8](#_Toc138599743)

[Power analysis 8](#_Toc138599744)

[Supplementary Figures 9](#_Toc138599745)

[References 11](#_Toc138599746)

## Supplementary Results

### Subthreshold genes

In order to investigate if there are any genes with strong associations with proxy AD/dementia signal which do not pass the Bonferroni correction threshold in the discovery dataset, we selected genes at Benjamini–Hochberg false-discovery rate (BH-FDR) of 10% in each of the 4 variant category SKAT-O analyses (**Supplementary Table 3.8**). No genes were identified in either of the pLOF categories, suggesting that the signal in these analyses was very sparse. Three additional genes were identified in the pLOF+REVEL>50 analysis; *FBXW4* (P=6.47x10-6, Nvariants=38), *BPIFB3* (P=1.80x10-5, Nvariants=13), *HS3ST2* (P=2.39x10-5, Nvariants=37). Removing singletons from these analyses does not considerably affect the results(*FBXW4*: P=3.93x10-6, Nvariants=20; *BPIFB3*: P=5.92x10-6, Nvariants=8; *HS3ST2*: P=2.25x10-5, Nvariants=11); however, removing all variants with MAC<5 does affect *BPIFB3* and *HS3ST2* considerably (*FBXW4*: P=4.93x10-6, Nvariants=10; *BPIFB3*: P=0.57, Nvariants=3; *HS3ST2*: P=0.65, Nvariants=1).

*FBXW4* encodes F-box/WD repeat-containing protein 4, a member of a family of proteins that contain F-box domains to recruit ubiquitin ligase and WD domains to recruit proteins to be labelled with ubiquitin for degradation1. *FBXW4* has been connected to acute myeloid leukemia1 but further disease associations are not well documented. Two other members of the F-box/WD family have been connected to AD through interaction with presenilin 1 (FBXW7)2 and a mouse model (*FBXW11*)3; however, the connection between *FBXW4* and AD is not well characterized. *BPIFB3* encodes BPI fold containing family B member 3, the function of this protein is not well characterized but it appears to impact infection response4,5. As with *FBXW4*, the connection between *BPIFB3* and AD is not clear. *HS3ST2* encodes heparan sulfate-glucosamine 3-sulfotransferase 2, an enzyme which adds 3-O-sulphated domains to heparin sulfates6. These 3-O-suphated domains have been hypothesized to contribute to AD by influencing the abnormal phosphorylation of tau6. Glycoproteins with heparan sulfates attached (Heparan sulfate proteoglycans) have been previously linked to many pathways that contribute to AD7,8. There appears to be some evidence for the role of *HS3ST2* in AD but further replication of the experimental work connecting *HS3ST2* to AD is needed. *FBXW4*, *BPIFB3*, and *HS3ST2* were not nominally significant in the replication dataset (**Supplementary Table 3.8**), which suggests that there is little evidence from rare variants to support these genes as AD/dementia genes.

Four additional genes were identified in the pLOF+missense analysis; *SORL1* (P=2.58x10-6; Nvariants=782), *HEXA* (P=3.82x10-6; Nvariants=196), *BFAR* (P=3.84x10-6; Nvariants=154), and *PVR* (P=2.31x10-5; Nvariants=158). *SORL1* and *HEXA* passed the Bonferroni correction threshold in the pLOF+REVEL>50 analysis and are discussed elsewhere. *PVR* is a gene in the *APOE* region and after conditioning on the *APOE* ε4 allele, *PVR* was not even nominally significant (P=0.36). This leaves *BFAR* as one interesting gene not previously highlighted in other analyses. *BFAR* was largely unaffected by the removal of singletons (*BFAR*: P=4.40x10-6, Nvariants=68) and variants with MAC<5 (*BFAR*: P=2.90x10-6, Nvariants=38). *BFAR* encodes for bifunctional apoptosis regulator (BAR), this protein sensitizes neurons to cell death if down regulated9. BAR binds to Bcl-210, a protein which inhibits apoptosis11 and has been implicated in AD12. It is possible that missense mutations in *BFAR* could limit the anti-apoptosis ability of BAR through reduced interaction with Bcl-2. However, *BFAR* was not nominally significant in the replication dataset (**Supplementary Table 3.8**), which suggests that there is little evidence from rare variants to support *BFAR* as an AD/dementia gene.

We then used all of the selected genes at BH-FDR rate of 10% in each of the 4 analyses to perform gene-set enrichment using FUMA GENE2FUNC3.This analysis identified 4 gene-sets significantly enriched with the selected genes after Bonferroni correction for the number of tested gene-sets. One MSigDB gene-set14 and three gene-sets based on GWAS catalog reported genes. The MSigSB gene-set based on genes which bind to low-density lipoprotein was significantly enriched (GOMF\_LOW\_DENSITY\_LIPOPROTEIN\_PARTICLE\_BINDING: PBonferroni= 0.045); however, only *SORL1* and *TREM2* overlapped with this gene-set. The GWAS catalog gene-sets were based on reported gene associations with AD phenotypes (Alzheimer's disease or family history of Alzheimer's disease: PBonferroni= 6.43x10-9; Alzheimer's disease: PBonferroni= 2.90x10-9; Alzheimer's disease (late onset): PBonferroni= 8.45x10-6). These results were driven by *TREM2*, *PVR*, *TOMM40*, and *SORL1*.

### Gene-set analyses

We aggregated the variants from previously identified genes (excluding genes significant in this study) into a gene-set and performed SKAT-O analyses across the four variant categories using this gene-set. Three of the gene-set analyses were nominally significant (HiC pLOF: *P*=0.0096, Nvariants=292; pLOF: *P*=0.013, Nvariants=362; pLOF+REVEL: *P*=0.0015, Nvariants=1653) with the analysis using pLOF and missense variants not being nominally significant (pLOF+missense: *P*=0.16, Nvariants=5467). The P-values of all of the gene-set analyses were lower than the median P-value of the set of genes included in the gene-set (HiC pLOF=0.23, pLOF=0.45, pLOF+REVEL=0.22, pLOF+missense=0.53). However, the gene-set P-values were approximately the same as the P-value of the most significant gene (*ABCA7*) included in that analysis, except from the pLOF+missense analysis which was considerably higher (*ABCA7*: HiC pLOF=0.02, pLOF=0.012, pLOF+REVEL=0.0017, pLOF+missense=0.0016). This suggests that the gene-set association signal in the nominally significant gene-set analyses may be mostly attributed to *ABCA7*. After removing *ABCA7* variants from the gene-set, none of the associations between the gene-set and proxy AD/dementia across the 4 variant categories were nominally significant (HiC pLOF: *P*=0.12, Nvariants=220; pLOF: *P*=0.59, Nvariants=281; pLOF+REVEL: *P*=0.14, Nvariants=1195; pLOF+missense: *P*=0.69, Nvariants=4431).

In order to identify biological processes implicated by rare variants associated with proxy AD/dementia, we performed gene-set analyses on 9769 gene-sets from MSigDB v7.014. Due to memory constraints we could only analyse gene-sets with <10,000 variants. We used the same SKAT-O model as with the gene analysis except all variants which mapped to the genes in the gene-set were aggregated together. After Bonferroni correction for 9769 gene-sets, we identified 33 significant gene-sets (**Supplementary Table 3.9**). All of these associations were robust to the exclusion of all variants in the larger *APOE* region (GRCh38: 19:40000000-50000000). However, none of these associations were robust to the exclusion of the variants in the larger *APOE* region (which includes *TOMM40*) and variants present in the 4 significant genes (*TREM2*, *HEXA*, *SORL1*, *TOMM40*). After removal of these variants only two of the significant gene-sets had even nominal significance. These two gene-sets (GO\_bp:go\_leukocyte\_activation\_involved\_in\_inflammatory\_response and GO\_bp:go\_astrocyte\_activation) each contained three genes (outside of the *APOE* region genes and *TREM2*, *HEXA*, *SORL1*, *TOMM40*) with nominal significance (P<0.05) in the pLOF+missense analyses. The three nominally significant genes in GO\_bp:go\_leukocyte\_activation\_involved\_in\_inflammatory\_response (*P*=0.034, Nvariants=7680) were *PTPRC* (*P*=0.0015, Nvariants=387), *MAPT* (*P*=0.018, Nvariants=275), and *GRN* (*P*=0.024, Nvariants=244). The three nominally significant genes in GO\_bp:go\_astrocyte\_activation (*P*=0.033, Nvariants=4722) were *MAPT* (*P*=0.018, Nvariants=275), *GRN* (*P*=0.024, Nvariants=244), *PSEN1* (*P*=0.040, Nvariants=113). This suggests that the associated gene-sets were almost entirely driven by variants in *TREM2*, *HEXA*, and *SORL1*, and that the gene-set analysis was not capturing additional information not found from the gene level analysis.

### *TOMM40* variants

*TOMM40* was significantly associated with proxy AD/dementia when all missense and pLOF variants were included in the variant aggregation analysis (*P*=2.39x10-10, Nvariants=119). There was one variant of interest in *TOMM40*; a significant rare missense variant (rs142412517; MAF=0.001731; *P*=4.43x10-10) **(Figure 3.2; Supplementary Table 3.4)**. This variant was also the only moderately associated variant in the replication dataset (*P*=5.05x10-12, MAF=0.00181). This missense variant (CADD=25.4; REVEL=0.238) causes an amino acid change from arginine to tryptophan (p.R239W (ENST00000252487.9)) in a conserved region, but the impact on the protein is unclear. This variant is in very low LD (<0.01) with the APOE ε4 variants (rs7412-C, rs429358-C), but when conditioning for APOE ε4 status based on these two variants, *TOMM40* was no longer significantly associated with the phenotype (*P*=0.043) **(Supplementary Table 3.2)**. No other significant gene lost significance when conditioned on APOE ε4 status. The low number of variants in the HiC pLOF, 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* ε4 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.

### *HEXA* variants

*HEXA* was significantly associated with the phenotype when restricting to pLOF and high confidence missense variants (pLOF+REVEL>50: *P*= 2.13x10-7, Nvariants=121). However, HEXA was not even nominally significant in the replication analysis. There was one moderately associated variant with a MAC>1 in *HEXA* in the discovery analysis, a high-confidence missense variant (rs121907970: MAF=0.003879, *P*=2.07x10-6) (**Supplementary Figure 3.2; Supplementary Table 3.4**). This variant is a missense variant (CADD=25.1; REVEL=0.925) which causes an amino acid substitution from arginine to tryptophan (p.R247W (ENST00000567159.5)/p.R258W (ENST00000566304.5)). 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 activity15.

In the discovery dataset, at least one risk allele of rs121907970 was present in 6 of the 277 (2.17%) individuals with an AD diagnosis, 14 of the 1075 (1.30%) individuals with two parents with an AD/dementia diagnosis, 195 of the 20,728 (0.94%) individuals with one parent with an AD/dementia diagnosis, and 933 of the 126,428 (0.74%) individuals with no AD diagnosis and no parent with an AD/dementia diagnosis. By ordering the outcome from least to most AD/dementia risk (no diagnosis and no affected parent → 1 affected parent → 2 affected parents → AD diagnosis), the individuals with higher AD/dementia risk were more likely to have at least one copy of the rs121907970 risk allele (two-sided ordinal chi-square/linear-by-linear association test: Z=4.278, P=1.89x10-5). However, rs121907970 was not even nominally significant in the replication dataset (*P*=0.46, MAF=0.0038) and the risk allele of this variant was not more likely to be present in individuals with higher AD/dementia risk (two-sided ordinal chi-square/linear-by-linear association test: Z=1.0089, P=0.313). This suggests that this variant is unlikely to be truly associated with AD/dementia.

*HEXA*, along with *HEXB*, encodes the protein β-hexosaminidase (Hex) which degrades GM2 gangliosides16. The accumulation of GM2 gangliosides leads to 3 rare neurodegenerative lysosomal storage disorders (Tay-Sachs, Sandhoff disease, and GM2 activator deficiency). GM2 accumulation is associated with neuronal structural change, microglial activation and response, and neuroglial dysfunction. There is also some evidence to support the colocalization of Hex with amyloid-beta plaques in the brains of AD patients17. However, given that *HEXA* and the sole highlighted variant were not even nominally significant in the replicated cohort, it is unlikely that HEXA is truly associated with proxy AD/dementia.

### Comparison to previously identified AD genes through common variants

We further looked at whether genes identified in common variant GWAS analyses were enriched in association signal. We selected genes highlighted in the ‘Known locus’ column of Table 1 and the ‘Gene’ column from Table 2 of Bellenguez *et al*. (2022)18 to act as AD genes previously associated through common variants. After excluding the genes implicated by the SKAT-O analyses (*APOE*, *TOMM40*, *SORL1*, *HEXA*, and *TREM2*) and the HLA region, 14 of the 73 AD genes had nominal significance (P<0.05) in at least one of the four SKAT-O analyses (**Supplementary Table 3.10**), including *ABI3* and *ABCA7* which were also implicated as AD genes through previous rare variant analyses. However, none of these genes reach significance after Bonferroni correction for the total number of common variant genes. To see if the association signal in these genes would combine in aggregate, we performed a gene-set analysis similar to the rare variant genes gene-set analysis described in the **Results** section of the main text, except using the genes identified through common variant GWAS. Due to memory limitations, we were unable to combine the variants in all of the common variant genes so we created two gene-sets, one for the genes in Table 1 (known loci) and one for the genes in Table 2 (new loci). The new loci gene-set was not nominally significant using any of the four variant categories (**Supplementary Table 3.10**); however, the known loci gene-set was nominally significant in the pLOF, pLOF+REVEL, and pLOF+missense variant category analyses (**Supplementary Table 3.10**). *ABCA7* was included in the known loci gene-set, so we tested whether these nominally significant results were robust to the exclusion of *ABCA7* variants and found that only the pLOF+missense category was nominally significant (P=0.0497) after removal of *ABCA7* variants. These results along with the results from the genes previously implicated by rare variants suggests that there is little association signal in previously implicated AD genes outside of *APOE* (*TOMM40*), *SORL1*, *TREM2*, and *ABCA7*.

### Association of known AD genes in true case-control analysis

We performed the SKAT-O variant aggregation analysis in the discovery dataset (as described in the main text) except using a case-control phenotype where individuals with an AD diagnosis were cases, and individuals without an AD diagnosis and without a parent with AD/dementia were controls (Ncases=277, Ncontrols= 126,428). While we identified, three known AD genes in these analyses (*ADAM10*, *BIN1*, *SORL1*), an additional 568 genes not previously associated with AD were also significant. The associations of *ADAM10*, *BIN1*, and *SORL1* were driven by one singleton that was present in cases. This suggests that the limited number of cases (277 cases) and variants in each gene allowed unreliable variants (singletons) to disproportionately affect the analysis leading to many false positives.

Additionally, we were interested in whether some known rare AD risk variants had higher allele frequencies in true AD cases compared to the individuals with the same phenotype value but derived from two affected parents rather than an AD diagnosis. We found that 66 variants across these genes were at least twice as frequent in true cases compared to individuals with two affected parents (**Supplementary Table 3.11**); however, 58 of those variants were singletons, and all of the variants had a minor allele count of five or lower in true cases. Only 10 of these variants have reported evidence for AD or dementia in ClinVar, and only two have been reported as pathogenic: rs63750900 in *PSEN1* (ENST00000324501.10: p.R269H) and rs770510230 in *ABCA7* (ENST00000263094.11: p.E1679X). This suggests that there was not a large difference in the frequency of rare AD associated variants in the true AD cases compared to the individuals with the same phenotype value but derived from parental case status. The proxy phenotype was representative of the associations of these genes in the true AD cases and was more reliable due to the inclusion of additional variants and individuals.

### *ABCA7* variants

We investigated the lack of association of *ABCA7* with the proxy phenotype by using the association of previously reported *ABCA7* variants in the discovery analysis and the variants with a moderate P-value (<1x10-4) in *ABCA7* in the discovery analysis. We identified variants reported in the studies19–24 highlighted as reporting rare variant associations in *ABCA7* with AD in Hoogmartens *et al.* (2021)25 (**Supplementary Table 3.12**). We connected these variants to variants in our study based on rsID, amino acid substitutions, genomic position, and alleles. We identified 53 previously highlighted *ABCA7* variants, 34 of which were only identified in studies of early onset AD. Of the total 53 previously highlighted variants, 23 were present in our discovery dataset, none of which had even a nominal association with proxy AD/dementia. In our discovery analysis, only one variant with a minor allele count >1 had a moderate P-value (rs545859049: tested allele=A; MAC=7; P=1.776x10-5), this variant is a missense variant that causes a substitution from arginine to glutamine on exon 20 (p.R907Q (ENST00000263094.11)) but is unlikely to affect the *ABCA7* protein (CADD=11.03, REVEL= 0.481).

### Power analysis

To identify the proportion of causal variants in a gene needed to have sufficient power to identify significant genes at the sample size used in the discovery analysis, we used SKAT-O power functions and modified the assumed number of causal variants in each gene. We identified that approximately 1.75-2% of variants in a gene need to be causal to have 80% power to observe significant genes at the Bonferroni corrected threshold after correcting for the number of genes and functional variant categories (P<6.78x10-7) (**Supplementary Table 3.13**). This analysis is based on the effective sample size calculated from the discovery dataset after dividing this number by 4 to correct for the proxy phenotype26(Neff=18,797). These results suggest that limiting a gene analysis to variants likely to have impacts on protein function will increase the power to observe a significant association. It also implies that, given the relatively small number of variants (100-300) used in the significant SKAT-O analyses, only a handful of causal variants were needed for the analysis to have sufficient power to identify these genes as significant.

## Supplementary Methods

### Subthreshold genes

We identified genes at Benjamini–Hochberg false-discovery rate (BH-FDR) of 10% in each of the 4 variant category SKAT-O analyses. One of these genes (*PVR*) was in the *APOE* region, so we repeated the SKAT-O analysis of this gene and included *APOE* ε4 status (based on rs7412-C and rs429358-C) as a covariate. We repeated the SKAT-O analyses for the BH-FDR genes without singletons and low MAC (<5) variants. Then using the full set of BH-FDR genes (N=9), including *TOMM40*, *HEXA*, *TREM2*, and *SORL1*, we performed gene-set enrichment using FUMA GENE2FUNC13 (v1.3.8). We compared the BH-FDR genes to the other genes tested in the analyses (N=18425); however, FUMA only recognized the Ensembl ID for 17,991 genes. The number of tests were adjusted using Bonferroni correction and the Bonferroni adjusted P-value cutoff was 0.05. Only gene-sets containing at least 2 overlapping genes were tested.

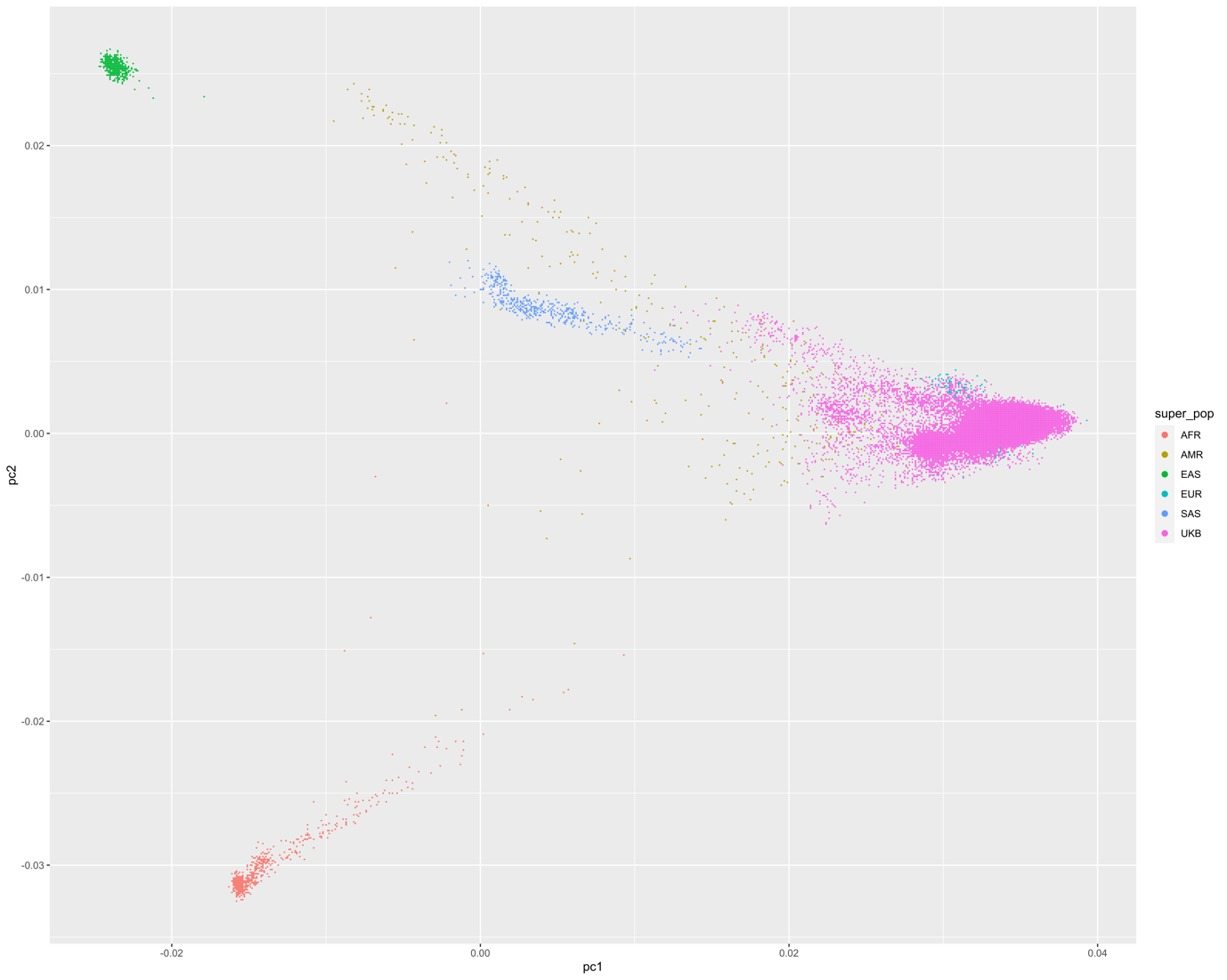
### SKAT-O gene-set analyses

We performed gene-set analyses to look if sub-threshold association signal would aggregate across genes to highlight biological processes. All pLOF and missense variants were used in the gene-set analysis because this was the variant category with the highest inflation and potentially the most association signal to be aggregated. We analysed 9769 gene-sets from MSigDB v7.014, we were only able to analyse gene-sets with <10,000 variants due to memory constraints. The gene-set analysis was performed using SKAT-O in the same way as the gene analyses, except all variants which mapped to the genes in the MSigDB gene-sets were aggregated together. We then repeated the analyses of the Bonferroni corrected significant gene-sets with variants in the larger *APOE* region (GRCh38: 19:40000000-50000000) removed. We again repeated the analyses with the variants in the larger *APOE* region and the variants in the significantly associated genes from the gene analysis (*TREM2*, *HEXA*, *SORL1*, *TOMM40*) removed.

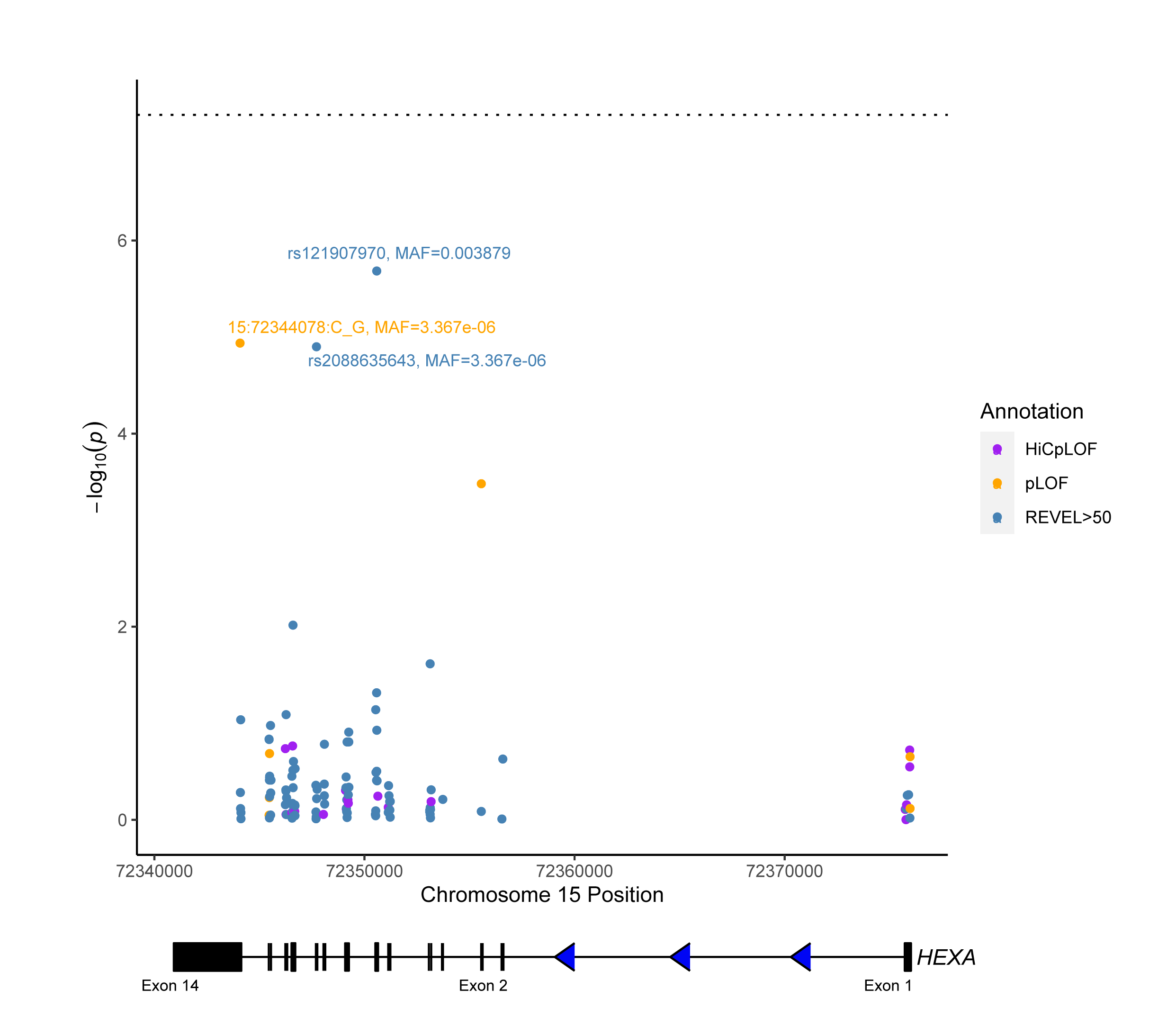
### Power analysis

The power analysis was implemented in the SKAT R package27 where the following parameters were assumed: the haplotype data was based on the data provided with the SKAT package, the causal MAF cutoff was 0.01, the alpha was the Bonferroni corrected P-value threshold (6.78x10-7), the sample size was the effective sample size of the discovery analysis divided by 4 to correct for the proxy phenotype26 (18,797), 100 simulations were performed for each causal variant percentage. We used both the continuous and binary phenotype version of the power functions and estimated power for the following causal variant percentages: 0.2, 0.5, 1, 1.75, 2, 2.5, 5, 10.

## Supplementary Figures



Supplementary Figure 3.1: The first and second principal component (pc1 and pc2, respectively) of the individuals included in the discovery and replication analyses and the 1000 Genomes reference population. Individuals not included in the UKB are coloured by their reported ancestry, individuals included in the UKB are coloured purple. Individuals were assigned to the continental ancestral superpopulation with the closest Mahalanobis distance based on principal components 1-30.



Supplementary Figure 3.2: The single variant associations of variants which mapped to HEXA. Each point represents a variant and is coloured based on functional annotation (HiCpLOF, pLOF, REVEL>50, missense). Missense variants were not plot for HEXA because the most significant result for this gene was in the pLOF+REVEL>50 analysis.

## References

1. Han, Q. *et al.* FBXW4 Is Highly Expressed and Associated With Poor Survival in Acute Myeloid Leukemia . *Frontiers in Oncology*  vol. 10 (2020).

2. Atkin, G. & Paulson, H. Ubiquitin pathways in neurodegenerative disease. *Front. Mol. Neurosci.* **7**, 63 (2014).

3. Sun, J. *et al.* FBXW11 deletion alleviates Alzheimer’s disease by reducing neuroinflammation and amyloid-β plaque formation via repression of ASK1 signaling. *Biochem. Biophys. Res. Commun.* **548**, 104–111 (2021).

4. Delorme-Axford, E. *et al.* BPIFB3 regulates autophagy and coxsackievirus B replication through a noncanonical pathway independent of the core initiation machinery. *MBio* **5**, e02147–e02147 (2014).

5. Evans, A. S., Lennemann, N. J. & Coyne, C. B. BPIFB3 Regulates Endoplasmic Reticulum Morphology To Facilitate Flavivirus Replication. *J. Virol.* **94**, e00029-20 (2020).

6. Sepulveda-Diaz, J. E. *et al.* HS3ST2 expression is critical for the abnormal phosphorylation of tau in Alzheimer’s disease-related tau pathology. *Brain* **138**, 1339–1354 (2015).

7. Snow, A. D., Cummings, J. A. & Lake, T. The Unifying Hypothesis of Alzheimer’s Disease: Heparan Sulfate Proteoglycans/Glycosaminoglycans Are Key as First Hypothesized Over 30 Years Ago. *Front. Aging Neurosci.* **13**, 710683 (2021).

8. Zhang, G., Zhang, X., Wang, X. & Li, J.-P. Towards understanding the roles of heparan sulfate proteoglycans in Alzheimer’s disease. *Biomed Res. Int.* **2014**, 516028 (2014).

9. Roth, W. *et al.* Bifunctional apoptosis inhibitor (BAR) protects neurons from diverse cell death pathways. *Cell Death Differ.* **10**, 1178–1187 (2003).

10. Zhang, H. *et al.* BAR: An apoptosis regulator at the intersection of caspases and Bcl-2 family proteins. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 2597–2602 (2000).

11. Kermer, P., Liman, J., Weishaupt, J. H. & Bähr, M. Neuronal Apoptosis in Neurodegenerative Diseases: From Basic Research to Clinical Application. *Neurodegener. Dis.* **1**, 9–19 (2004).

12. Callens, M. *et al.* The role of Bcl-2 proteins in modulating neuronal Ca2+ signaling in health and in Alzheimer’s disease. *Biochim. Biophys. Acta - Mol. Cell Res.* **1868**, 118997 (2021).

13. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).

14. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).

15. 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).

16. Toro, C., Zainab, M. & Tifft, C. J. The GM2 gangliosidoses: Unlocking the mysteries of pathogenesis and treatment. *Neurosci. Lett.* **764**, 136195 (2021).

17. 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).

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

19. Cuyvers, E. *et al.* Mutations in <em>ABCA7</em> in a Belgian cohort of Alzheimer’s disease patients: a targeted resequencing study. *Lancet Neurol.* **14**, 814–822 (2015).

20. De Roeck, A. *et al.* Deleterious ABCA7 mutations and transcript rescue mechanisms in early onset Alzheimer’s disease. *Acta Neuropathol.* **134**, 475–487 (2017).

21. Le Guennec, K. *et al.* ABCA7 rare variants and Alzheimer disease risk. *Neurology* **86**, 2134–2137 (2016).

22. Logue, M. W. *et al.* Targeted Sequencing of Alzheimer Disease Genes in African Americans Implicates Novel Risk Variants. *Front. Neurosci.* **12**, 592 (2018).

23. Steinberg, S. *et al.* Loss-of-function variants in ABCA7 confer risk of Alzheimer’s disease. *Nat. Genet.* **47**, 445–447 (2015).

24. 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).

25. 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).

26. 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).

27. 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).