scientific reports

Check for updates

OPEN

Novel Alzheimer's disease genes and epistasis identified using machine learning GWAS platform

Mischa Lundberg[1,2,3,10]✉, Letitia M. F. Sng[1,10], Piotr Szul[4], Rob Dunne[5], Arash Bayat[6], Samantha C. Burnham[7], Denis C. Bauer[1,8,9,11] & Natalie A. Twine[1,9,11]✉

Alzheimer's disease (AD) is a complex genetic disease, and variants identified through genome-wide association studies (GWAS) explain only part of its heritability. Epistasis has been proposed as a major contributor to this 'missing heritability', however, many current methods are limited to only modelling additive effects. We use VariantSpark, a machine learning approach to GWAS, and BitEpi, a tool for epistasis detection, to identify AD associated variants and interactions across two independent cohorts, ADNI and UK Biobank. By incorporating significant epistatic interactions, we captured 10.41% more phenotypic variance than logistic regression (LR). We validate the well-established AD loci, *APOE*, and identify two novel genome-wide significant AD associated loci in both cohorts, *SH3BP4* and *SASH1*, which are also in significant epistatic interactions with *APOE*. We show that the *SH3BP4* SNP has a modulating effect on the known pathogenic *APOE* SNP, demonstrating a possible protective mechanism against AD. *SASH1* is involved in a triplet interaction with pathogenic *APOE* SNP and *ACOT11*, where the *SASH1* SNP lowered the pathogenic interaction effect between *ACOT11* and *APOE*. Finally, we demonstrate that VariantSpark detects disease associations with 80% fewer controls than LR, unlocking discoveries in well annotated but smaller cohorts.

Alzheimer's disease (AD) is the most common form of dementia and predominantly affects individuals over 65[1]. The vast majority (99%) of AD cases are late onset (LOAD) and are driven by multiple genetic and environmental influences, with genetics accounting for between 53 and 80% of total phenotypic variance[2–4]. The heritability of LOAD is predominantly carried by the *APOE* locus, which explains about 25% of the total heritability of the disease[5]. In addition to *APOE*, large-scale genome-wide association study (GWAS) meta-analyses identified 40[6] and 75 additional risk loci[7], but more than 30% of genetic variability remains unknown[3]. Recent studies[8,9] predict that there are 100 to 1000 causal variants with modest effects associated with LOAD, of which only a small proportion have been identified.

Part of the missing heritability in LOAD might be explained by non-additive interactions[10], which are ignored by GWAS studies. Indeed, a genome-wide replicated scan has found epistasis to be a ubiquitous phenomenon across multiple phenotypes[11]. Epistatic interactions have long been implicated in complex genetic disease, including neurological diseases[12] and LOAD itself[13]. However, due to the computational complexity of finding genome-wide gene–gene interactions, the search were limited to candidate gene approaches[13–17], or genome-wide approaches exploring interactions between *APOE* and other risk loci[18].

Using the ML platform VariantSpark[19], we overcome the shortcomings of traditional statistical GWAS approaches and computationally limited epistatis discovery tools to identify genome-wide variants associated with LOAD and AD in both the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort (512 cases, 272 controls)[20] and the UK Biobank (UKBB) cohort (704 cases, up to 6869 controls)[21]. Using a novel false discovery

[1]Transformational Bioinformatics, Commonwealth Scientific and Industrial Research Organisation, Sydney, NSW, Australia. [2]UQ Frazer Institute, The University of Queensland, Woolloongabba, QLD, Australia. [3]Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD, Australia. [4]Health Data Semantics and Interoperability, Commonwealth Scientific and Industrial Research Organisation AU, Brisbane, QLD, Australia. [5]Data61, Commonwealth Scientific and Industrial Research Organisation, Brisbane, QLD, Australia. [6]The Kinghorn Cancer Center (KCCG), Garvan Institute of Medical Research, Sydney, NSW, Australia. [7]Biomedical Imaging Group, CSIRO, Brisbane, QLD, Australia. [8]Department of Biomedical Sciences, Faculty of Medicine and Health Science, Macquarie University, Macquarie Park, NSW, Australia. [9]Applied BioSciences, Faculty of Science and Engineering, Macquarie University, Macquarie Park, NSW, Australia. [10]These authors contributed equally: Mischa Lundberg and Letitia M. F. Sng. [11]These authors jointly supervised this work: Denis C. Bauer and Natalie A. Twine. ✉email: m.lundberg@uq.net.au; natalie.twine@csiro.au

rate (FDR) method[22], we are able to use VariantSpark's random-forest-based feature selection approach to narrow down the genome-wide search space to the subset of variants enriched with epistatic interactions. We then apply BitEpi[23] to perform an exhaustive search of this subset to annotate pairwise and higher-order, statistically significant interactions between the variants. We also explore the proportion of phenotypic variance captured by VariantSpark versus the traditional logistic regression (LR) methods. Finally, we demonstrate that VariantSpark has improved sensitivity to detect signal with fewer control samples compared with LR approaches.

Results

VariantSpark identifies known AD loci across two independent cohorts

Using the ML genomics platform VariantSpark[19], and a novel RFlocalfdr approach[22], we identified genetic variants that are both marginally and interactively associated in two independent AD cohorts, UKBB and ADNI (7,573 and 784 samples of 4.5M SNPs each). Because of these two types of associations, we expect to find more significant variants than a LR approach at < 5% FDR.

We identified 104 SNPs (53 independent) to be significantly associated with AD in the UKBB cohort (Table 1, Fig. 1, Supplementary Table S1) and 207 significantly associated SNPs (124 independent) in the ADNI cohort (Fig. 1, Supplementary Table S2). When we compared these associations with those associated with AD in the GWAS Catalog (trait ID 'MONDO_0004975, accessed 16/05/22)[24] using locus bins, we observed a 70% overlap with the significant SNPs identified in both the UKBB (72/104) and ADNI cohort (145/207), with 31 out of the 53 independent UKBB SNPs (58.49%) and 82/124 (66.13%) of the independent ADNI SNPs (Table 1 and Supplementary Table 2).

As expected, the *APOE* loci was identified in both ADNI and UKBB cohorts (Supplementary Tables 1 and 2). To evaluate the functional context of the other significantly associated independent variants, we performed functional enrichment analysis using MAGMA. Gene-set analysis (Supplementary Table S4) identified 9 (ADNI) and 3 (UKBB) gene sets significantly associated (after Bonferroni correction). Many of the significant gene sets and those with suggestive significance levels ($P < 0.05$) fell into the categories of transmembrane and metal ion transport proteins (known to be key in neuronal signalling in the brain). Tissue expression analysis using MAGMA and GTEX (Supplementary Table S5, Supplementary Table S6) revealed brain tissues to be the most highly ranked, although they did not pass Bonferroni correction.

VariantSpark identifies novel loci associated with AD

We next investigated which loci replicated between the two independent cohorts. Despite the phenotypic heterogeneity across the two cohorts, we replicated three independent, significantly associated genes, *APOE* (rs429358), *SASH1*, and *SH3BP4* (Table 1 and Supplementary Table S2). It is important to note that the significance threshold for the RFlocalfdr is 0.05 compared to the traditional genome-wide significance threshold of $P < 5 \times 10^{-8}$, which needs to correct for multiple tests. Both thresholds, RFlocalfdr for VariantSpark and $P < 5 \times 10^{-8}$ for logistic regression, control for Type 1 error and correct for the multiple testing burden. For further information, see Methods section.

Both *SASH1* and *SH3BP4* were novel to our study and were not yet present in the GWAS Catalog SNPs, although there is a marginally associated SNP (rs9390537, χ^2-$p = 8.17 \times 10^{-6}$) mapping to an intergenic region 91,233 bp upstream of *SASH1* associated with AD[25] and another marginally associated SNP (rs66501349, χ^2-$p = 2 \times 10^{-6}$) intergenic to *SH3BP4* and *CEP19P1* associated with poorer cognitive function[26]. The corresponding rsIDs from the UKBB cohort are rs117160741 (Chr 6:148512131) for *SASH1* and rs114656810 (Chr 2:235751287) for *SH3BP4* (Supplementary Table S1). Both are intergenic and located upstream of the genes. Similarly, the rsIDs from the ADNI cohort are rs9918382 (Chr 6:148265029), an intergenic variant located upstream of *SASH1,* while rs6711272 (Chr 2:235131361) is an intergenic variant located downstream of *SH3BP4* (Supplementary Table S2).

SASH1 (SAM and SH3 domain-containing 1) encodes a scaffold protein, which is ubiquitously expressed, including in brain tissues and is also a positive regulator of the NF-$_k$B signalling pathway through the activation of *TLR4*[27]. *SH3BP4* (SH3 domain binding protein 4) encodes a protein involved in the amino acid-induced TOR signalling pathway[28]. Both *SASH1* and *SH3BP4* are membrane bound phosphoproteins with SH3 domains.

BitEpi identifies novel interactions between known and novel AD genes

BitEpi was used to identify epistatic interactions between significantly associated variants in both cohorts. The β and α metrics, reflecting association power and interaction effect respectively, were used to select interactions that were strongly associated to the AD phenotype due to an epistatic effect. We identified 37 interactions with significant β and α values in the UKBB cohort, of which 17 were 2-SNP, 16 were 3-SNP, and 4 were 4-SNP interactions (Fig. 2, Supplementary Table S7). Using the ADNI cohort, we identified 58 interactions with significant β and α values, 39 were 2-SNP, 17 were 3-SNP and 2 were 4-SNP interactions (Fig. 3, Supplementary Table S8). Interestingly, the two replicating AD associated genes, *SASH1* and *SH3BP4*, were involved in epistatic interactions.

In the UKBB cohort, the SNP (rs114656810) mapping to *SH3BP4* was found to interact with rs429358, which is a reported pathogenic *APOE* SNP in ClinVar[29], where the alternate 'C' allele plays a part in the high AD-risk APOE-ε4 isoform. This pairwise interaction was interrogated to identify the genotype combinations associated with AD (Supplementary Table S9). Due to the low number of samples with the homozygous alternate genotype (AA) of *SH3BP4* SNP, we reduced the genotypes to two classes; presence or absence of the alternate 'A' allele. In the absence of the alternate *SH3BP4* SNP allele, there was no absolute difference in control rates between the *SH3BP4*x*APOE* interaction and the *APOE* SNP alone (Fig. 4A). This indicates a limited effect of the homozygous reference genotype of rs114656810 on AD. However, with the presence of the alternate allele of the *SH3BP4* SNP, the pathogenic effect of the *APOE* C allele is modulated (Fig. 4A), suggesting that *SH3BP4* may have a protective

Chr	Pos	Alt	RSID	*P* value	Gene	ADNI	GWAS catalog
1	174049377	C	rs72711440	4.36E+09	RABGAP1L-DT	FALSE	FALSE
1	193137934	C	rs139963893	2.50E+09	CDC73	FALSE	TRUE
1	244007787	A	rs75965920	5.16E+09	AKT3	FALSE	FALSE
2	38520113	A	rs56302953	3.05E+09	ATL2	FALSE	TRUE
2	205966824	G	rs116192932	1.64E+09	PARD3B	FALSE	TRUE
2	215673948	C	rs71579843	1.70E+09	BARD1	FALSE	FALSE
2	235751287	G	rs114656810	2.29E+09	SH3BP4	TRUE	FALSE
3	76534878	C	rs74801337	4.85E+09	ROBO2	FALSE	FALSE
3	154755590	C	rs9829241	1.39E+09	MME	FALSE	FALSE
4	116269757	A	rs147896092	6.36E+05	NDST4	FALSE	FALSE
4	178201067	G	rs76175875	2.34E+08	NEIL3	FALSE	TRUE
5	36648950	C	rs115172522	1.07E+09	SLC1A3	FALSE	TRUE
5	93761812	G	rs112217406	2.73E+09	KIAA0825	FALSE	FALSE
5	153560523	G	rs71585927	3.90E+09	GALNT10	FALSE	TRUE
5	168782459	C	rs78269616	1.07E+07	SLIT3	FALSE	FALSE
6	46454250	G	rs114866534	1.49E+09	RCAN2	FALSE	TRUE
6	81120137	G	rs117103821	1.75E+07	BCKDHB	FALSE	TRUE
6	133182939	NA	rs57823471	5.59E+09	NA	TRUE	TRUE
6	148512131	C	rs117160741	5.15E+07	SASH1	TRUE	TRUE
6	150085133	A	rs117243801	2.84E+08	PCMT1	FALSE	TRUE
7	101930310	C	rs11552019	5.76E+09	SH2B2	FALSE	TRUE
8	52718273	T	rs112585504	3.10E+09	PXDNL	FALSE	FALSE
8	65042594	T	rs78789176	5.43E+08	LINC01414	FALSE	TRUE
9	25640336	A	rs77057081	1.16E+09	TUSC1	FALSE	TRUE
9	108857477	C	rs117992330	3.40E+09	TMEM38B	FALSE	TRUE
9	136144593	G	rs66697526	3.65E+08	ABO	FALSE	TRUE
10	66463703	C	rs189699806	4.56E+07	ANXA2P3	FALSE	FALSE
10	70366297	C	rs61868095	3.59E+08	TET1	FALSE	TRUE
10	122310126	C	rs79486209	5.24E+09	PLPP4	FALSE	FALSE
11	12566857	C	rs76154502	7.12E+08	PARVA	FALSE	TRUE
11	44909137	C	rs78150932	2.09E+07	TSPAN18	FALSE	TRUE
11	118613605	A	rs139648410	1.08E+09	DDX6	FALSE	FALSE
12	65040623	G	rs79774071	1.12E+09	RASSF3	FALSE	TRUE
12	107592329	G	rs75378184	9.57E+08	CRY1	FALSE	TRUE
13	22104766	T	rs116942699	4.31E+09	MICU2	FALSE	FALSE
13	55121441	G	rs117658626	5.37E+07	MIR1297	FALSE	FALSE
14	62969759	G	rs79473324	1.39E+09	KCNH5	FALSE	FALSE
14	99261321	C	rs142983586	4.35E+07	C14orf177	FALSE	TRUE
16	53388460	A	rs2908783	2.29E+09	LOC643802	FALSE	FALSE
17	2761430	C	rs74252350	2.63E+09	RAP1GAP2	FALSE	FALSE
17	54490063	C	rs147441895	4.16E+08	ANKFN1	FALSE	FALSE
18	22582278	T	rs79110874	3.25E+09	LINC01894	FALSE	FALSE
18	77685936	G	rs113702893	4.25E+09	SLC66A2	FALSE	FALSE
19	17262529	C	rs117951200	1.12E+09	MYO9B	FALSE	TRUE
19	45411941	T	rs429358	0.00E+00	APOE	TRUE	TRUE
19	55511927	G	rs1560714	2.98E+09	NLRP2	FALSE	TRUE
21	36714721	A	rs2834914	1.22E+09	LOC100506403	FALSE	FALSE

Table 1. Annotated statistically significant independent SNPs identified using VariantSpark and the UKBB cohort. Replication status to ADNI validation cohort and GWAS Catalog included.

mechanism against AD for carriers of the *APOE* 'CC' genotype. In the ADNI cohort, this pairwise interaction between *SH3BP4* and *APOE* was marginally significant but did not pass Bonferroni correction.

In the ADNI cohort, the SNP rs9918382 mapping to *SASH1* was involved in a triplet interaction with the same pathogenic APOE SNP, rs429358. The other SNP, rs7552961, in the triplet maps to *ACOT11*, has been shown to be associated to mild cognitive decline[30]. This triplet interaction was also examined further (Supplementary Table S10). Again, due to the low numbers of samples with the homozygous alternate genotype of rs9918382

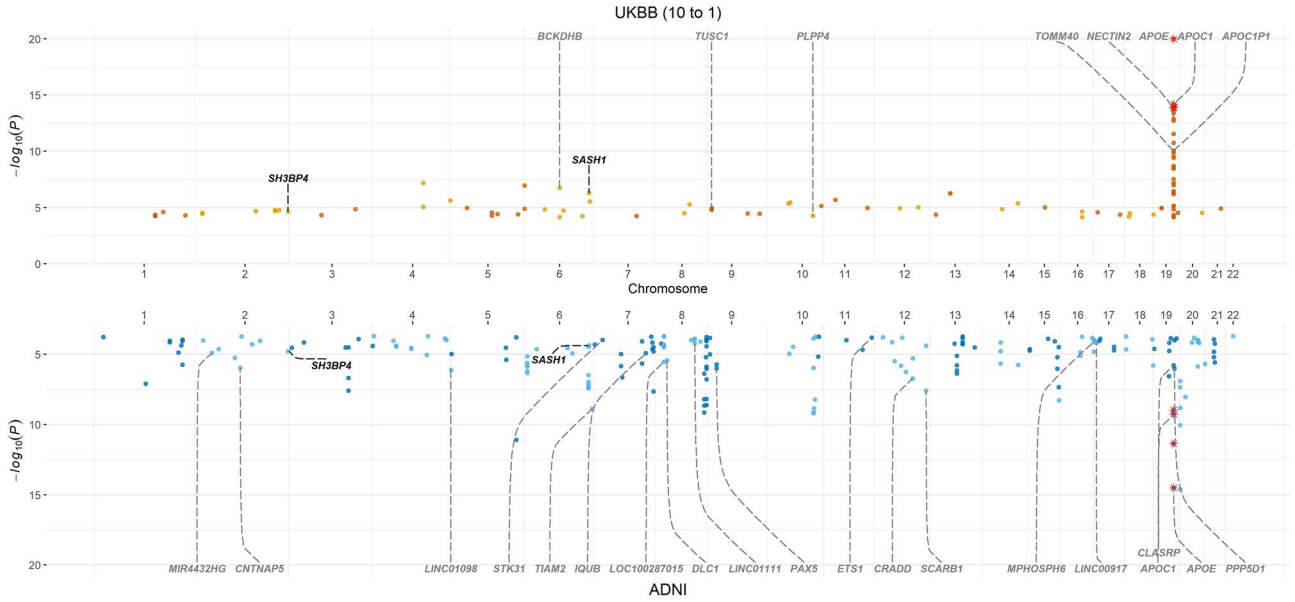

Figure 1. Miami plot showing significant SNPs identified by VariantSpark in UK Biobank (10 controls to 1 case) (top) and ADNI (bottom) cohorts. Red asterisks mark those variants that have been replicated by position (only independent variants) between the two cohorts. Annotation (black) represents gene annotations that are novel and replicated between the two cohorts. Annotations in grey represent previously identified variants.

($n = 15$), the genotype was reduced to two classes; presence or absence of the alternate 'G' allele. Figure 4B shows that the alternate 'G' allele of the *SASH1* SNP has a protective effect, reversing the pathogenic interaction effect of the rs7552961 (*ACOT11*) TT genotype and rs429358 (*APOE*) TC genotype increasing the relative control rate from −0.139 to 0.028 (Supplementary Table S10). However, when the alternate *ACOT11* allele (G) is present with the *APOE* CC genotype, the *SASH1* SNP has no effect. In fact, none of the possible pairwise interactions between these three genotypes passed significance for the α metric, which suggests that the association to AD was carried by the interaction of all three SNPs. This highlights the complexity and difficulty of detecting epistatic interactions, where exacerbating or protective properties are exerted through specific combinations of genotypes.

VariantSpark can detect more disease associated signal than logistic regression

Next, we compared VariantSpark with the more traditional GWAS approach implemented in PLINK's logistic regression (LR) to estimate the power to detect disease associated signal with limited control samples. To do this, in addition to using the ADNI cohort, we subset two datasets from the UKBB cohort: the first contained a ratio of 10 controls to 1 case (UKBB10to1) and the second with 2 controls to 1 case (UKBB2to1).

Using LR, we did identify multiple variants at suggestive significance levels using the ADNI cohort (ranging from χ^2-$p = 8.34 \times 10^{-8}$ to χ^2-$p = 2.63 \times 10^{-6}$), all falling into the *APOE* locus (Chr19:45,326,217 to Chr19:45,445,517). Based on the UKBB cohort, we identified three significantly independent associated SNPs in UKBB10to1 (127 in total) (Supplementary Table S3) and one significantly independent associated SNP in UKBB2to1 (74 in total) (Supplementary Table S11). All SNPs found using LR fell within the *APOE* locus (Chr19:45,326,217 to Chr19:45,445,517).

In contrast, VariantSpark identified associations outside of the *APOE* region such as rs79486209 on chromosome 10 which mapped to *PLPP4*, a gene previously associated with AD[31]. VariantSpark identified 53 significantly associated independent SNPs (104 in total) in UKBB10to1 (Table 1) and 20 significantly associated independent SNPs (69 in total) in UKBB2to1 (Supplementary Table S12).

This demonstrates we have 15% (1/3 vs. 20/53) more power to detect disease associated variants with 80% fewer (2 vs. 10) controls using VariantSpark compared with a LR approach.

VariantSpark captures more phenotypic variance in AD than Logistic Regression

A key goal of this study was to explore whether epistasis can explain some of the missing heritability that is well documented in AD[2–4]. To this end, we measured the proportion of phenotypic variance captured by genetic variants identified in the UKBB cohort using Nagelkerke's pseudo-R^2 and fitting three LR models with: Firstly, significant and independent SNPs identified by LR (n = 3). Secondly, significant and independent SNPs identified by VariantSpark (n = 53). Thirdly, significant and independent SNPs identified by VariantSpark with significant interactions identified by BitEpi (n = 122).

Within the UKBB cohort, the VariantSpark-BitEpi model (model (3)) captured the highest variance explained at 23.18% compared to model (2) without the BitEpi interactions at 17.12% and model (1) the LR SNPs at 12.77% (Supplementary Fig. S2). To test whether the performance increase of the VariantSpark-BitEpi model was driven by its additional variables, we calculated an empirical *P* value. We fitted 1000 models containing the 3 LR SNPs

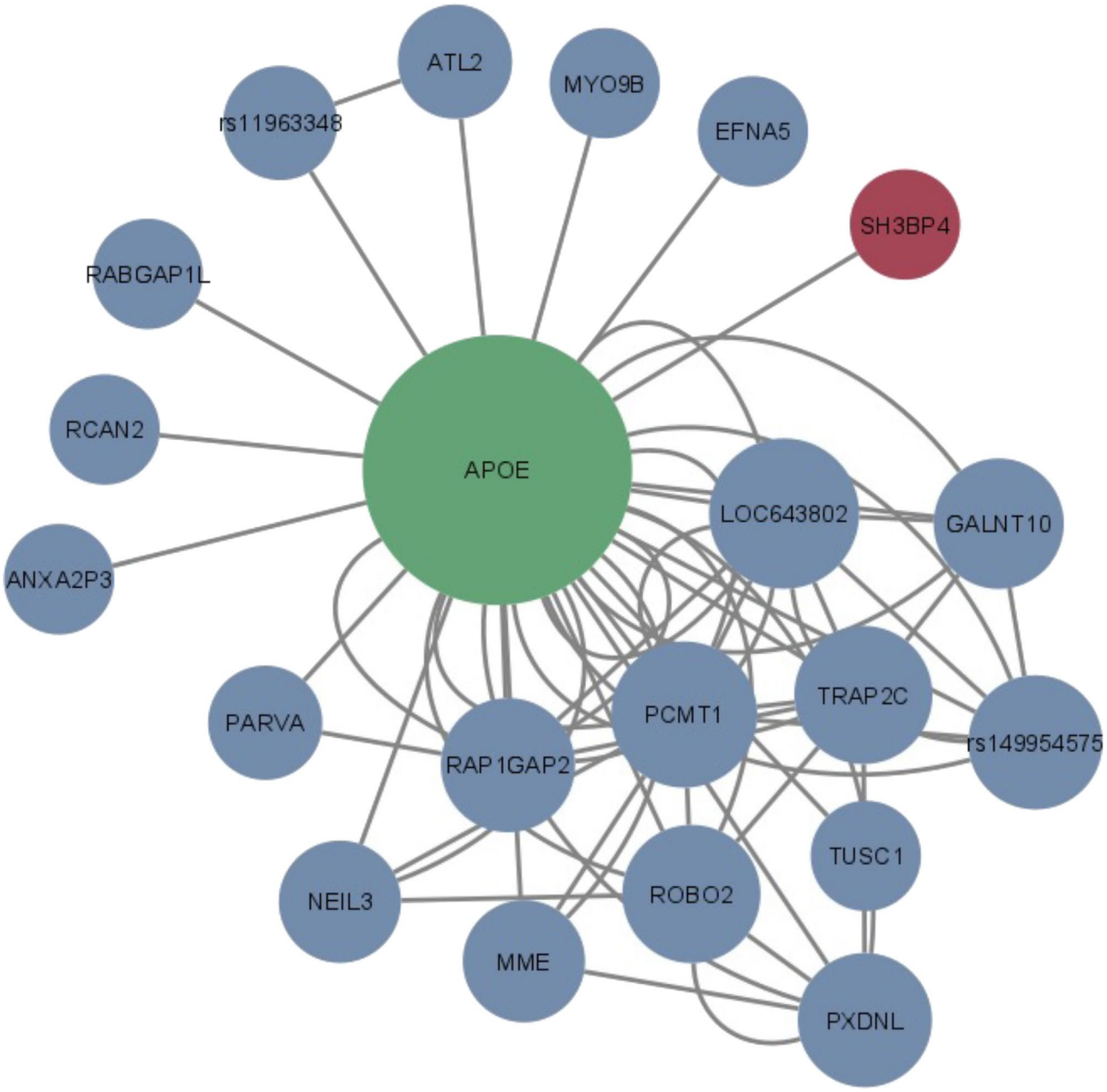

Figure 2. Network diagram of significant BitEpi interactions from UK Biobank cohort. Nodes in green are known AD associations, in red is the novel gene replicated in this study, and in blue are variants which are novel but unreplicated. All 2-SNP, 3-SNP, and 4-SNP interactions are included. Size of nodes are representative of node degree calculated from the NetworkAnalyzer plug-in in Cytoscape.

as well as 50 randomly selected SNPs and 69 interactions to emulate the degrees of freedom of the VariantSpark-BitEpi model (3). As shown in Supplemental Fig. S2, these models achieved an average pseudo-R^2 of 19.33%, outperforming the models with fewer predictors (models (1) and (2)). In contrast, VariantSpark-BitEpi's model had a small but significant ($p = 0.006$) performance improvement over the random models (23.18% vs 19.33%), confirming that additional signal was captured. We make a similar observation for these models when tested on the independent ADNI cohort. LR (model 1) captured 7.09% while the random models captured 25% on average and VariantSpark-BitEpi (model 3) achieved 27.20%. The increase in variance explained on the ADNI set is likely due to an easier signal, which is predominantly driven by *APOE* (as observed in Section C).

These findings indicate that VariantSpark-identified SNPs and BitEpi-identified epistatic interactions together explain up to 10.41% more phenotypic variance in AD than traditional LR approaches that focus only on marginal effects. This also aligns with previous studies where the addition of 87 marginal effect SNPs (without *APOE*) explained only 2.1% more variance[32] and 2,042,105 SNPs (without known AD SNPs) accounted for 25.3% variance[3]. Taken together, these results suggest that epistatic interactions across the genome play a part in AD aetiology and should be accounted for when developing therapeutics and genetic risk scores.

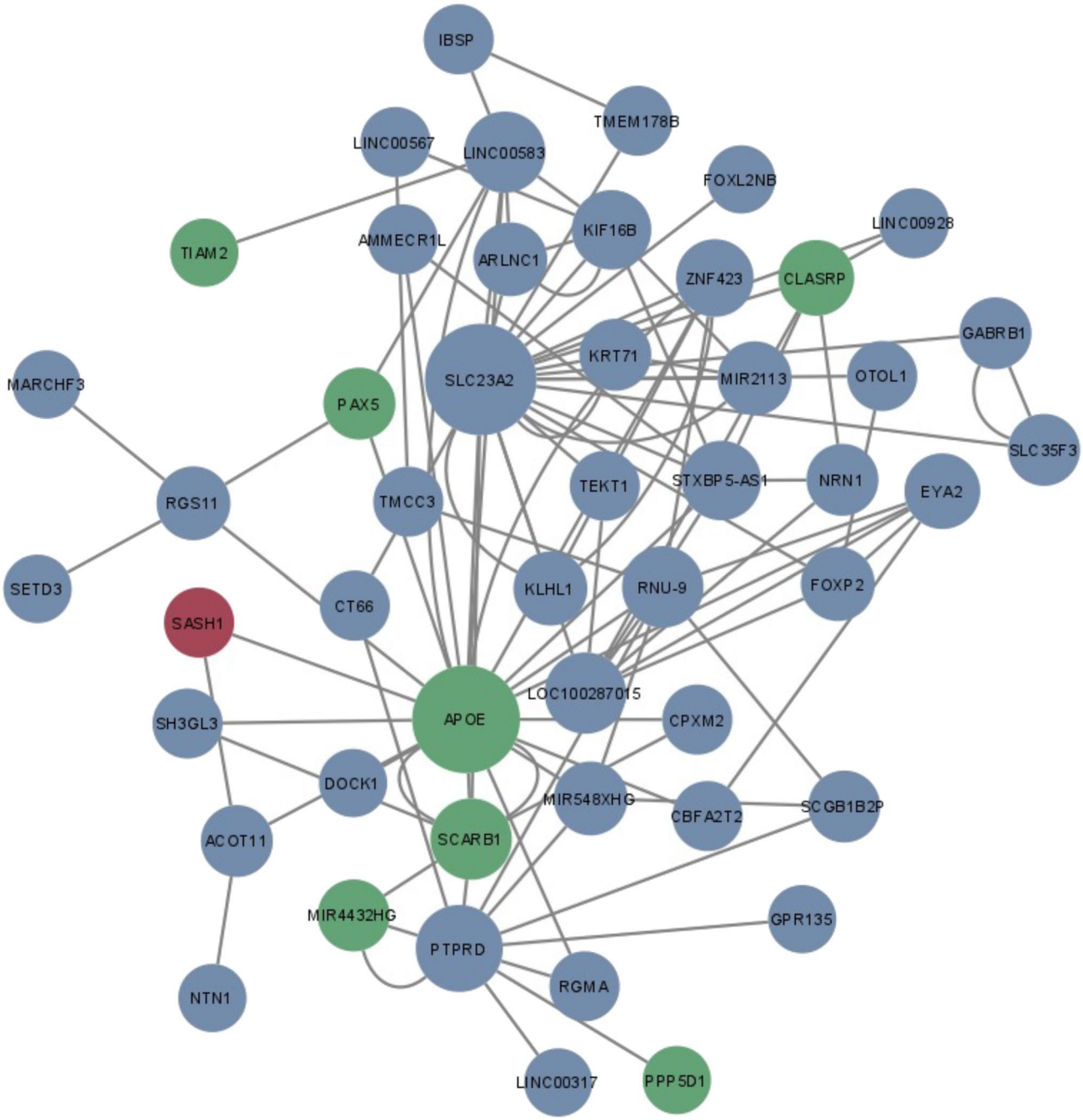

Figure 3. Network diagram of significant BitEpi interactions from ADNI cohort. Nodes in green are known AD associations, in red is the novel gene replicated in this study, and in blue are variants which are novel but unreplicated. All 2-SNP, 3-SNP, and 4-SNP interactions are included. Size of nodes are representative of node degree calculated from the NetworkAnalyzer plug-in in Cytoscape.

Transcriptome-wide association (TWAS) lookup of SASH1 and SH3BP4

Finally, we looked at transcriptomic level information of the mapped genes *SASH1* and *SH3BP4* as in previous studies[33,34] have shown that this can add confidence that GWAS-identified genes are capturing actual disease-related signal. Using the TWAS-hub[35], *SASH1* showed strong evidence (ENET-$P = 7.5 \times 10^{-9}$) of involvement in the prefrontal cortex tissue and a strong association with "Alzheimer's Disease (in father)" (Supplementary Table S13). In contrast, *SH3BP4* showed an association with nerve tibial tissue at non-suggestive levels for Alzheimer's Disease (Supplementary Table S14). Another resource used were the gene expression tests built into FUMA[36] using GTEx v8[37] data. In this analysis, both *SASH1* and *SH3BP4* showed increased expression levels in brain tissue (Supplementary Fig S3).

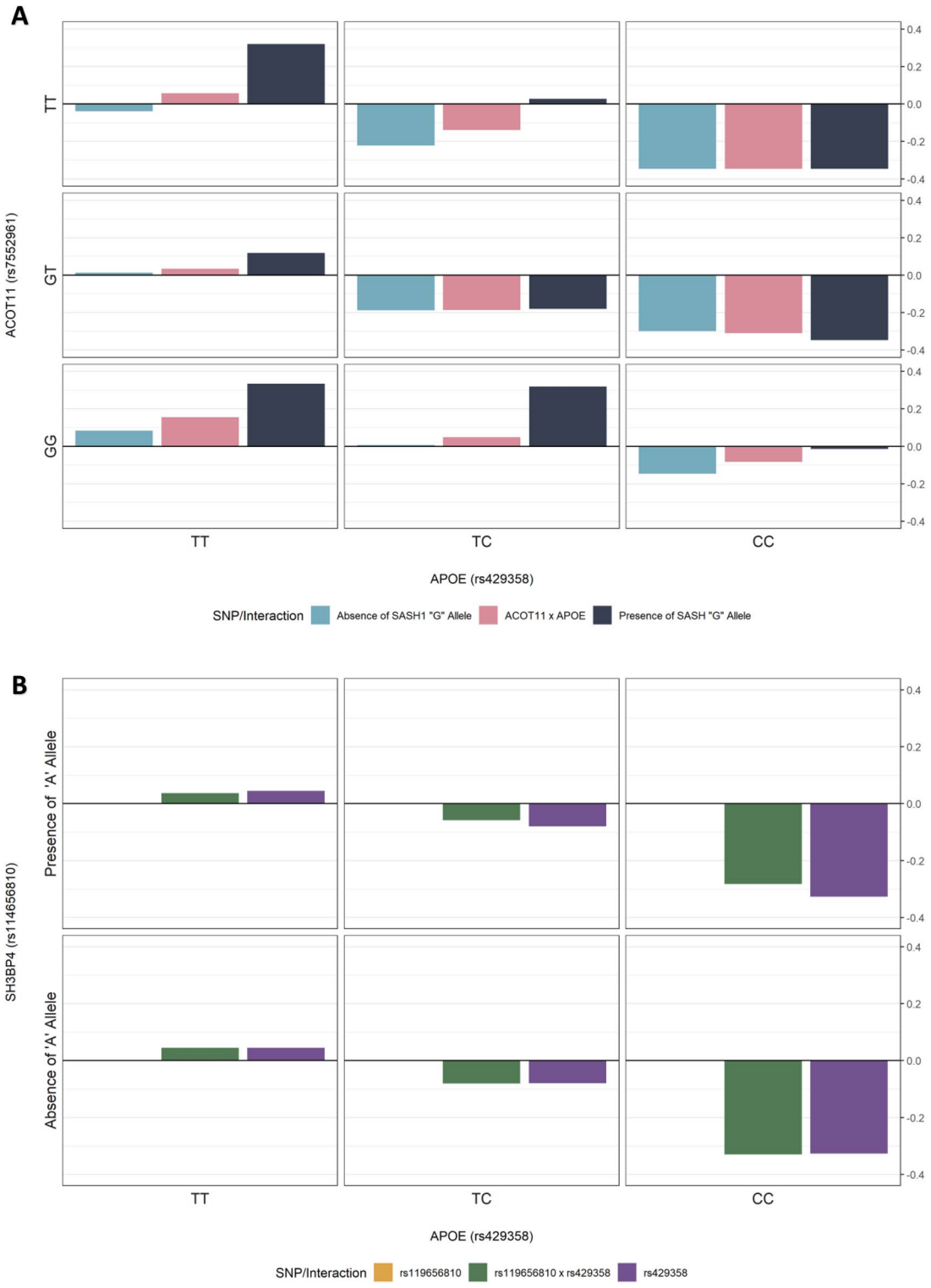

Figure 4. Relative control rates of the interactions (**A**) rs119656810 (*SH3BP4*) and rs429358 (*APOE*) in the UK Biobank cohort, and (**B**) rs7552961 (*ACOT11*), rs9918382 (*SASH1*), and rs429358 (*APOE*) in the ADNI cohort. Relative control rates were calculated as the difference between control rates of each genotype combination and the control rate of the entire cohort. Due to sample size restrictions, the rs119656810 SNP and the rs9918382 SNP was reduced to two categories; presence or absence of its alternate allele. There is evidence of a modulating effect of the alternate allele of rs119656810 on the *APOE-e4* (rs429358 CC) genotype as seen from the increase in relative control rates in the top middle and top right cells in (**A**). There is evidence of a protective effect of alternate allele of rs9918382 on the *ACOT11* × *APOE* genotypes as seen from the increase in relative control rates in the top middle cell and the bottom right cell in (**B**). However, there is no evidence of the same effect for the *APOE-e4* (rs429358 CC) genotype in an interaction with the *ACOT11* alternate allele (rs7552961).

Discussion

Using VariantSpark, a ML approach to GWAS, we have identified two novel genes, *SASH1* and *SH3BP4,* to be associated with AD reaching genome-wide significance.

SASH1 is a known tumour suppressor protein that has been shown to be differentially expressed between AD and control samples[38,39]. Furthermore, a previous study found SNP rs9390537 (located 91,233 bp upstream of *SASH1)* to be nominally associated to LOAD (χ^2-$p = 8.17 \times 10^{-6}$)[25]. Indeed, it is a nominated AD drug target in the Agora database, a database curated by AD researchers from the accelerating medicine partnership-Alzheimer's disease consortium and other research teams.

SH3BP4 or transferrin trafficking protein (*TTP*) interacts with endocytic proteins including clathrin, dynamin, and the transferrin receptor[40] and is involved in the aminal acid-Rag GTPase-mTORC1 signalling pathway. It is a central link between Akt signalling and cell–matrix adhesion regulation[28]. Although *SH3BP4* has no established link to AD, a SNP (rs66501349, intergenic to *SH3BP4* and *CEP19P1*) has been marginally associated to poorer cognitive function (χ^2-$p = 2 \times 10^{-6}$)[26] and its interactor dynamin has strong evidence of a role in AD pathophysiology[41,42]. In particular, the expression of gene *DNM2* was significantly decreased in AD patients, and neuronal cell lines transfected with dominant negative *DNM* genes were observed to have an accumulation of APP and increased Aβ secretion[43].

The key contribution of our work is adding the lens of epistasis to association. We identified a total of 95 epistatic interactions, including 2-SNP, 3-SNP and 4-SNP interactions associated with AD, in two independent cohorts. This elevated the previously only nominally associated *SASH1*[25] to pass FDR significance when its interaction with *ACOT11* and *APOE* is accounted for. Specifically, our epistasis analysis revealed that the alternate 'G' allele of *SASH1* SNP rs9918382 appears to have a protective effect against AD as it reverses the pathogenic effect of *ACTO11* rs7552961 'TT' and *APOE* rs429358 'TC' genotype combination (Supplementary Fig. S3). However, this modulating effect was not found in the presence of two copies of the pathogenic *APOE* 'C' allele (rs429358, Supplementary Fig. S3). This result is consistent with co-expression patterns found between AD and control brains[44] and the high expression levels of *SASH1* in pre-frontal cortex tissue in the TWAS-hub. Taken together, it is likely that *SASH1* plays a role in AD pathophysiology and warrants further investigations.

Although, most of our identified epistasis is concentrated between *APOE* and a small number of other loci, our methodology can explore genome-wide epistasis in an unbiased manner, unlike previous studies[45,46]. Additionally, a genome-wide search allows for the identification of epistasis in non-coding regions of the genome which have empirically demonstrated to effect gene expression[47].

For example, our epistasis analysis revealed a modulating effect of the alternate allele of SNP *rs*119656810 (*SH3BP4*) on the *APOE* locus. A possible explanation for this effect is that *SH3BP4* has the ability to regulate the activity of dynamin[40], whereby it enables the processing of amyloid β protein precursors resulting in lower levels of Aβ depositions and AD pathology. Together, *SH3BP4* is a novel gene that may play a role in AD pathophysiology through its pathway mechanisms and in combination with *APOE*.

While VariantSpark identified *SH3BP4* and *SASH1* in both cohorts due to their cumulative additive and epistatic effects on AD, the exact epistatic interactions they are involved in were not replicated, although *SH3BP4-APOE* showed marginal significance. This is likely due to the varying number of individuals who might have this exact modulating disease physiology and genotype combinations across the two cohorts. This illustrates the benefits of using VariantSpark instead of traditional LR models on binary traits with potential polygenic interactions, like Alzheimer's disease.

Using VariantSpark, we were also able to detect disease genes with fewer controls than traditional approaches. This is relevant as a recent study calculates 10,000,000 cases would be needed for a traditional GWAS to find significant SNPs explaining 50% of Alzheimer's disease heritability[48]. Even for large initiatives such as FinnGen or 23andMe, such numbers are hard to achieve. Our method offers an alternative and enables discoveries in smaller but well annotated cohorts for AD and other genetic studies.

The limitations to our study are as follows: Firstly, ADNI used whole genome sequencing mapped to the GRCh38 reference genome, while the UKBB used array technology mapped to the GRCh37 reference genome resulting in the final set of 4.5 million common SNPs which was around 50% of the total number of SNPs for both cohorts. Secondly, the ADNI and UKBB cohorts are both different ascertainments. Particularly, UKBB is a relatively healthy volunteer cohort and contained a mix of AD phenotypes while ADNI recruited patients based on their health status and included samples with mild cognitive impairment to maximise sample size but is only an AD-proxy phenotype. Lastly, the ADNI cohort was substantially smaller than the UKBB cohort, with 784 samples, compared to 7582 samples. These three factors in combination limit our power to discover and replicate disease variants and epistatic interactions across cohorts. Furthermore, we restricted our samples in both cohorts to those of European descent as is commonplace[49]. However, it has been shown that ethnicity plays a crucial role in AD aetiology[50,51] and more diverse genomic datasets are needed to gain better unbiased insights[52].

In conclusion, we have established a ML approach for detecting genetic signals associated with disease, which goes some way to explain the missing heritability observed in previous literature.

Methods

Sample selection

Data for AD was obtained from two sources; ADNI and UKBB. The ADNI aimed at testing combinations of imaging and biological markers to measure progression of AD and mild cognitive impairment (MCI). For this study, cases were samples labelled as early and late MCI and AD (Supplementary Note 1). The UKBB contains phenotypic and biological information from 500,000 participants; see their previous publication for more details[21]. For this study, ICD10 codes from hospital inpatient records and participant responses were used to identify cases of AD. See supplementary for specific codes, question, and responses used. Additionally, individuals with indication

of early onset AD and/or family history of AD were excluded. Based on the UK Biobank two subsets were generated to identify differences in detection power for novel variants. One contained a ratio of 1 case to 2 controls (labelled UKBB2to1) and the other a ratio of 1 case to 10 controls (labelled UKBB10to1). The UKBB10to1 cohort was used for all result sections, unless specified. Counts of individuals included in the analyses are shown in Supplementary Note 1. This research was approved by the UK Biobank's governing Research Ethics Committee.

Quality control

Quality control (QC) included exclusion of variants with minor allele frequency (MAF) < 0.01, imputation quality < 0.9, genotype missingness > 0.1 and those deviating from Hardy–Weinberg equilibrium ($P < 1 \times 10^{-6}$). Furthermore, individuals with a discrepancy between their genetic and reported sex were excluded and if their genotype-derived principal components 1 and 2 were further than 6 standard deviations away from those of 1000 Genomes European population. After QC, we had 11.7 M variants in UKBB, and 9.5 M variants in ADNI, with 4.6 M in common between the two cohorts. Notably, the ADNI cohort was mapped to the GRCh38 reference while the UKBiobank was mapped to the GRCh37 reference.

Genome-wide association study using logistic regression

Association testing between AD and genetic variants was conducted using whole genome LR model implemented in PLINK[53] (v1.90beta). Sex, age and the top 20 principal components were used as covariates for the association analysis.

Genome-wide association study using VariantSpark

VariantSpark[19], a distributed implementation of the random forest (RF) algorithm, was used for association testing on Amazon Web Services. The same QC'd input files from LR analyses were used in the VariantSpark analyses. Optimisation of four hyperparameters; mTry, minNodeSize, MaxDepth, and nTree was run on all cohorts The optimised settings for all three cohorts were the same for mTry (0.1), MaxDepth (10), and nTree (20,000) except for minNodeSize where UKBB10to1 = 758, UKBB2to1 = 211, and ADNI = 78.

We determined the reliability of VariantSpark on real datasets by comparing Gini importance score of three runs on the UKBB10to1 and ADNI cohorts as Pearson's correlations (Supplementary Figs. S1). Further, we tested the effect of covariates (as used in LR) in a RF model by comparing the out-of-bag error metric between a Ranger[54] run with covariates and a VariantSpark run without covariates. We did not observe any difference between the models; thus, covariates were not included in the final VariantSpark analysis.

Compute resources

LR analyses were conducted on a machine with 16 Cores and 48 GiB memory. VariantSpark analyses were conducted using AWS Elastic Map Reduce with a total sum of 64 vCores and 488 GiB of memory.

Post-GWAS analyses

P value calculation

The primary measure of association from VariantSpark is the importance score derived from Gini-Index[55]. While this score can rank variants by importance, it is unable to determine significantly associated variants. To determine significance from importance scores, we used a recently developed method[22]. Briefly, this approach is based on the empirical Bayes method[56] which uses RF tree information as a threshold to fit a skew normal distribution and correct for multiple testing akin to Efron's local false discovery rate approach.

Identification of independent variants, functional mapping and annotation

Variants identified in the GWAS were annotated using SNPTracker[57] and clumped using PLINK v.1.90b3.31[53] within a window of 1000 kb and r^2 of 0.01. Significantly associated variants were functionally mapped and annotated using ANNOVAR (v.7 2020-06-08)[58]. Furthermore, all significantly associated variants were mapped into locus bins where each locus bin was created based on a two million base-pair sliding window around the variants. This allowed known associations from the GWAS Catalog to be mapped to our results by identifying bins that are shared between the GWAS Catalog and our study's associations.

General quality assurance of the UKBB (discovery) and ADNI (replication) cohort

PLINK LR results were used to identify potential population stratification using LDSC. No evidence for inflated statistics due to hidden population stratification was detected (LDSC intercept estimate was 1.03 ± 0.01 and 1.03 ± 0.01 for UKBB10to1 and ADNI, respectively).

Epistasis calculation using BitEpi

To identify 2-SNP, 3-SNP, and 4-SNP interactions, BitEpi was applied to the significant VariantSpark associations in the UKBB and ADNI cohorts separately. The methods behind BitEpi have already been discussed elsewhere[59] but briefly, BitEpi calculates two entropy metrics, α and β. The β metric reflects the combined association power of all the SNPs involved in the interaction while the α metric represents the gain in association power due to the epistatic effect of all interactive SNPs. Therefore, an interaction with a large α and β has a strong association with the phenotype caused by an epistatic effect between all of the SNPs in the interaction. Quantiles for each order (2-SNP, 3-SNP or 4-SNP interactions) were used to filter out interactions with higher α and β values before *P*-values were computed through a permutation procedure. Bonferroni-corrected significance thresholds were

calculated based on all possible combinations, with < 0.05 denoting significance. SNPs involved in significant interactions were annotated with their independent SNP to remove any redundant interactions.

Using an in-house Python script, we generated contingency tables for some of the significant interactions found by BitEpi (Fig. 4, Supplementary Table S9). The control rate is the number of controls over the number of samples for each genotype combination or for the overall cohort. The relative control rate is then the overall control rate minus the genotype combination control rate. A genotype combination with a negative relative control ratio can be considered to be deleterious and vice versa.

Variance explained calculation

The significant associations from the VariantSpark, PLINK LR, and BitEpi analyses using the UKBB cohort were used to calculate the variance explained calculated as Nagelkerke's pseudo-R2[60] within the UKBB and ADNI cohort with the following as predictors in logistic models run using R v4.1.3[61]; (1) significant and independent VariantSpark SNPs ($n = 53$), (2) significant and independent PLINK LR SNPs ($n = 3$), (3) significant and independent VariantSpark SNPs and all significant BitEpi interactions as interacting variables ($n = 122$). For all three models, the response was the AD case/control status. An empirical *P*-value was calculated from 1000 'random noise' models which were built to mimic the structure of model 3 by including the known *APOE* SNPs found by VariantSpark but also SNPs with no association with AD.

Data availability

The data that support the findings of this study are available from ADNI database (https://adni.loni.usc.edu/data-samples/access-data/) and through the UK Biobank Data Showcase (http://www.ukbiobank.ac.uk/).

Received: 4 April 2023; Accepted: 7 October 2023
Published online: 17 October 2023

References

1. Winblad, B. *et al.* Defeating Alzheimer's disease and other dementias: A priority for European science and society. *Lancet Neurol.* **15**, 455–532 (2016).
2. Gatz, M. *et al.* Heritability for Alzheimer's disease: The study of dementia in Swedish twins. *J. Gerontol. A Biol. Sci. Med. Sci.* **52**, M117-125 (1997).
3. Ridge, P. G. *et al.* Assessment of the genetic variance of late-onset Alzheimer's disease. *Neurobiol. Aging* **41**(200), e13-200.e20 (2016).
4. So, H.-C., Gui, A. H. S., Cherny, S. S. & Sham, P. C. Evaluating the heritability explained by known susceptibility variants: A survey of ten complex diseases. *Genet. Epidemiol.* **35**, 310–317 (2011).
5. Van Cauwenberghe, C., Van Broeckhoven, C. & Sleegers, K. The genetic landscape of Alzheimer disease: Clinical implications and perspectives. *Genet. Med.* **18**, 421–430 (2016).
6. Andrews, S. J., Fulton-Howard, B. & Goate, A. Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. *Lancet Neurol.* **19**, 326–335 (2020).
7. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).
8. Holland, D. *et al.* Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLoS Genet.* **16**, e1008612 (2020).
9. Zhang, Q. *et al.* Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. *Nat. Commun.* **11**, 4799 (2020).
10. Mackay, T. F. & Moore, J. H. Why epistasis is important for tackling complex human disease genetics. *Genome Med.* **6**, 124 (2014).
11. Chatelain, C. *et al. Atlas of epistasis*. (Genetic and Genomic Medicine, 2021). https://doi.org/10.1101/2021.03.17.21253794.
12. Sha, Q., Zhang, Z., Schymick, J. C., Traynor, B. J. & Zhang, S. Genome-wide association reveals three SNPs associated with sporadic amyotrophic lateral sclerosis through a two-locus analysis. *BMC Med. Genet.* **10**, 86 (2009).
13. Hohman, T. J. *et al.* Discovery of gene–gene interactions across multiple independent data sets of late onset Alzheimer disease from the Alzheimer Disease Genetics Consortium. *Neurobiol. Aging* **38**, 141–150 (2016).
14. Arosio, B. *et al.* Interleukin-10 and interleukin-6 gene polymorphisms as risk factors for Alzheimer's disease. *Neurobiol. Aging* **25**, 1009–1015 (2004).
15. Heun, R. *et al.* Interactions between PPAR-α and inflammation-related cytokine genes on the development of Alzheimer's disease, observed by the Epistasis Project. *Int. J. Mol. Epidemiol. Genet.* **3**, 39–47 (2012).
16. Kauwe, J. S. K. *et al.* Suggestive synergy between genetic variants in TF and HFE as risk factors for Alzheimer's disease. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153B**, 955–959 (2010).
17. Mateo, I. *et al.* Interaction between dopamine beta-hydroxylase and interleukin genes increases Alzheimer's disease risk. *J. Neurol. Neurosurg. Psychiatry* **77**, 278–279 (2006).
18. Belbin, O. *et al.* Investigation of 15 of the top candidate genes for late-onset Alzheimer's disease. *Hum. Genet.* **129**, 273–282 (2011).
19. Bayat, A. *et al.* VariantSpark: Cloud-based machine learning for association study of complex phenotype and large-scale genomic data. *Gigascience* **9**, giaa007 (2020).
20. Petersen, R. C. *et al.* Alzheimer's disease neuroimaging initiative (ADNI): Clinical characterization. *Neurology* **74**, 201–209 (2010).
21. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
22. Dunne, R. Threshold values for the gini variable importance a empirical bayes approach. 2022.04.06.487300 Preprint at https://doi.org/10.1101/2022.04.06.487300v1 (2022).
23. Bayat, A. *et al.* BitEpi: A fast and accurate exhaustive higher-order epistasis search. *bioRxiv* 858282.https://doi.org/10.1101/858282 (2020).
24. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
25. Heinzen, E. L. *et al.* Genome-wide scan of copy number variation in late-onset Alzheimer's disease. *J. Alzheimer's Dis.* **19**, 69–77 (2010).
26. Homann, J. *et al.* Genome-wide association study of Alzheimer's disease brain imaging biomarkers and neuropsychological phenotypes in the European medical information framework for Alzheimer's disease multimodal biomarker discovery dataset. *Front. Aging Neurosci.* **14**, 840651 (2022).
27. Jaufmann, J. *et al.* The emerging and diverse roles of the SLy/SASH1-protein family in health and disease—Overview of three multifunctional proteins. *FASEB J* **35**, e21470 (2021).

28. Burckhardt, C. J., Minna, J. D. & Danuser, G. SH3BP4 promotes neuropilin-1 and α5-integrin endocytosis and is inhibited by Akt. *Dev. Cell* **56**, 1164-1181.e12 (2021).
29. National Center for Biotechnology Information. ClinVar; [VCV000017864.16]. https://www.ncbi.nlm.nih.gov/clinvar/variation/17864/?new_evidence=false
30. Hu, X. *et al.* Genome-wide association study identifies multiple novel loci associated with disease progression in subjects with mild cognitive impairment. *Transl. Psychiatry* **1**, e54–e54 (2011).
31. Sherva, R. *et al.* Genome-wide association study of the rate of cognitive decline in Alzheimer's disease. *Alzheimers Dement.* **10**, 45–52 (2014).
32. Karlsson, I. K. *et al.* Measuring heritable contributions to Alzheimer's disease: Polygenic risk score analysis with twins. *Brain Commun.* **4**, fcab308 (2022).
33. Cao, C. *et al.* Power analysis of transcriptome-wide association study: Implications for practical protocol choice. *PLoS Genet.* **17**, e1009405 (2021).
34. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
35. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245–252 (2016).
36. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
37. THE Gtex CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
38. Lv, L., Zhang, D., Hua, P. & Yang, S. The glial-specific hypermethylated 3′ untranslated region of histone deacetylase 1 may modulates several signal pathways in Alzheimer's disease. *Life Sci.* **265**, 118760 (2021).
39. Tan, M. G. *et al.* Genome wide profiling of altered gene expression in the neocortex of Alzheimer's disease. *J. Neurosci. Res.* **88**, 1157–1169 (2010).
40. Tosoni, D. *et al.* TTP specifically regulates the internalization of the transferrin receptor. *Cell* **123**, 875–888 (2005).
41. Cao, Y., Xiao, Y., Ravid, R. & Guan, Z.-Z. Changed clathrin regulatory proteins in the brains of Alzheimer's disease patients and animal models. *J Alzheimers Dis* **22**, 329–342 (2010).
42. Wu, F. & Yao, P. J. Clathrin-mediated endocytosis and Alzheimer's disease: An update. *Ageing Res. Rev.* **8**, 147–149 (2009).
43. Kamagata, E. *et al.* Decrease of dynamin 2 levels in late-onset Alzheimer's disease alters Aβ metabolism. *Biochem. Biophys. Res. Commun.* **379**, 691–695 (2009).
44. Narayanan, M. *et al.* Common dysregulation network in the human prefrontal cortex underlies two neurodegenerative diseases. *Mol. Syst. Biol.* **10**, 743 (2014).
45. Wang, H., Bennett, D. A., De Jager, P. L., Zhang, Q.-Y. & Zhang, H.-Y. Genome-wide epistasis analysis for Alzheimer's disease and implications for genetic risk prediction. *Alzheimers Res. Ther.* **13**, 55 (2021).
46. Gusareva, E. S. *et al.* Genome-wide association interaction analysis for Alzheimer's disease. *Neurobiol. Aging* **35**, 2436–2443 (2014).
47. van de Haar, J. *et al.* Identifying epistasis in cancer genomes: A delicate affair. *Cell* **177**, 1375–1383 (2019).
48. O'Connor, L. J. The distribution of common-variant effect sizes. *Nat. Genet.* **53**, 1243–1249 (2021).
49. Lleó, A. & Suárez-Calvet, M. Race and Alzheimer disease biomarkers. *Neurol. Genet.* **7**, e574 (2021).
50. Rubin, L. *et al.* Genetic risk factors for Alzheimer's disease in racial/ethnic minority populations in the U.S.: A scoping review. *Front. Public Health* **9**, 784958 (2021).
51. Schindler, S. E. *et al.* African Americans have differences in CSF soluble TREM2 and associated genetic variants. *Neurol. Genet.* **7**, e571 (2021).
52. Mills, M. C. & Rahal, C. The GWAS diversity monitor tracks diversity by disease in real time. *Nat. Genet.* **52**, 242–243 (2020).
53. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4,** s13742-015-0047-8 (2015).
54. Wright, M. N. & Ziegler, A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).
55. Nembrini, S., König, I. R. & Wright, M. N. The revival of the Gini importance?. *Bioinformatics* **34**, 3711–3718 (2018).
56. Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Cambridge University Press, Cambridge, 2010). https://doi.org/10.1017/CBO9780511761362.
57. Deng, J.-E., Sham, P. C. & Li, M.-X. SNPTracker: A Swift tool for comprehensive tracking and unifying dbSNP rs IDs and genomic coordinates of massive sequence variants. *G3 (Bethesda)* **6**, 205–207 (2015).
58. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
59. Bayat, A. *et al.* Fast and accurate exhaustive higher-order epistasis search with BitEpi. *Sci. Rep.* **11**, 15923 (2021).
60. Nagelkerke, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692 (1991).
61. R Core Team. *R: A Language and Environment for Statistical Computing* (2021). https://www.R-project.org/.

Acknowledgements

Funding was from CSIRO Health and Biosecurity. The results published here are in whole or in part based on data obtained from Agora, a platform initially developed by the NIA-funded AMP-AD consortium that shares evidence in support of AD target discovery. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Samantha Burnham is now employed at "Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly and

Company." Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Author contributions

N.A.T, D.C.B, M.L conceptualised the study. L.M.F.S, P.S, R.D, M.L, A.B, N.A.T, D.C.B contributed to primary data analysis, including statistical analysis. N.A.T, D.C.B, M.L, L.M.F.S wrote the manuscript. S.C.B contributed to the interpretation of the data. All authors critically reviewed and gave final approval for the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-44378-y.

Correspondence and requests for materials should be addressed to M.L. or N.A.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.