Brookes KJ, Farrell C. Utilising Polygenic Risk Score Analysis for AD to Determine the "Sphere of Influence" of the APOE Isoform SNPs. J Neurol Neuromedicine (2022) 6(2): 1-7

www.jneurology.com

J Neurology Neuromedicine
www.jneurology.com
Journal of Neurology & Neuromedicine

Original Research Article Open Access

Utilising Polygenic Risk Score Analysis for AD to Determine the "Sphere of Influence" of the APOE Isoform SNPs

Connor Farrell, Keeley J Brookes*

Biosciences, Nottingham Trent University, Clifton Campus, Nottingham, NG8 11NS, UK

Article Info

Article Notes
Received: May 30, 2022
Accepted: July 26, 2022

***Correspondence:**
*Dr. Keeley J Brookes, Biosciences, Nottingham Trent University, Clifton Campus, Nottingham, NG8 11NS, UK; Email: keeley.brookes@ntu.ac.uk.

Keywords:
Alzheimer's disease
Polygenic risk score
Independent association
Linkage disequilibrium
NECTIN2
TOMM40
APOC1
APOE

ABSTRACT

The *APOE* gene and particularly the ε4 allele have been a long-established risk factor for Alzheimer's disease (AD), demonstrating the largest genetic effect size in this complex disease. In light of the odds ratios observed for the risk allele, many studies disregard neighbouring association signals as merely "tagging" this effect. Polygenic risk score (PRS) analyses in this field regularly use low linkage disequilibrium parameters ($r^2 \geq 0.1$) when selecting SNPs for analysis across the genome and remove kilobases of data surrounding the *APOE* locus, preventing confounding factors influencing their results. This study investigated a 500kb region surrounding the *APOE* locus, utilising PRS analysis to explore whether additional SNPs in this region could be providing contributory effects to AD predictability. The data presented here suggest that the "sphere of influence" of the *APOE* isoform SNPs covers a region of around 92kb; SNPs in Linkage Disequilibrium (LD) at $r^2 < 0.4$ with rs429358 potentially contribute independently to the PRS predictability for AD, and that there are additional independent SNPs in this region that have increased effects in an *APOE* ε4 negative sample. This study concludes that further consideration is required when selecting LD parameters for PRS analysis and that additional investigation into the region surrounding *APOE* may yield polymorphisms that may play a pivotal role in the development of AD.

Introduction

The Apolipoprotein E (*APOE*) gene is singularly the most replicated genetic association with Alzheimer's disease (AD), located on chr19q13, its potential role has been highlighted since the early 1990's through linkage studies[1]. The *APOE* ε4 allele/isoform is determined by the genotypes present at two coding SNPs, rs429358 and rs7412, and has been observed to have a multiplicative dosage effect on AD risk, with odds ratio estimated at 3 for heterozygous carriers and up to 12 for homozygotes[2,3]. Despite its large effect size, this genetic variant does not account for the entire estimated heritability observed for AD[4], and therefore genetic variants have been sought and found in genome-wide studies over the past 30 years[5]. In recent years, association analyses have given way to polygenic risk score (PRS) studies, summarizing the effect sizes of multiple alleles into a single score with the aim to differentiate between cases and controls[6]. These studies have highlighted the involvement of multiple polymorphisms in the AD phenotype; however, these studies often exclude the *APOE* region. This exclusion is due to observations of multiple association signals in neighbouring SNPs to rs429358 and rs7412, believed to be tagging the effect of the isoform SNPs via linkage disequilibrium (LD), therefore large regions of the genome surrounding *APOE* are often removed prior to PRS analysis to prevent confounding effects of these variants. This excluded region ranges from 14kb to over 2Mb[7–12], despite multiple studies suggesting that additional loci in this

region may be having independent effects[13–19]. Therefore, the removal of this area could be missing key contributory variants from PRS models. This study aimed to investigate the extent of the "sphere of influence" of the *APOE* isoform SNPs by exploring the clumping parameters within a 500kb region, removed in our previous studies[20], and identify potential SNPs within this region that may be independent contributors to the polygenic risk score for AD.

Methods

Datasets

The IGAP_stage 1 (IGAP_S1) summary statistics[21] were used as the base dataset, and genotype data from the Brains for Dementia Research (BDR) project[22] was used as the target dataset to generate the polygenic risk scores. The BDR dataset underwent standard quality control with PLINK v1.9[23]. removing SNPs with a minor allele frequency of less than 1%; genotype calls of less than 95% and which deviated significantly from Hardy-Weinberg Equilibrium (P<0.0001) in the control samples. Samples that had less than 95% call rate were also removed. This resulted in an analysis dataset of 520 samples consisting of 356 pathological confirmed AD cases, and 164 controls.

Clumping

The IGAP_S1 summary statistics were clumped using the 1000Genomes dataset in PLINK v1.9[23], using the parameters –clump-p1 1 –clump-p2 1 –clump-kb 250 and –clump-r2 ranging from 0.1 to 0.9. Using R v4.0.3[24] the clumped output file for each r2 was changed from wide-to-long format for comparison with the .bim files from the target datasets, allowing each common SNP to be tagged with a clump number identifier.

SNPs in a 500kb region (hg19_chr19:45,160,844-45,660,844) surrounding the *APOE* isoform SNPs were extracted from both datasets, and common SNPs present in both datasets were carried forward into the analysis.

Polygenic Risk Score Generation

Polygenic risk scoring was carried out using the –score parameter in PLINK v1.9[23]. Logistic regression was carried out in R v4.0.3[24], followed by calculating the Area Under the Curve (AUC) using the pROC package[25]. Using R v4.0.3 a 'magic for loop' [26] was set up to allow the inclusion of SNPs from the base dataset individually across the *APOE* region (R script available upon request).

Results

Sixty-two SNPs were present in both the IGAP_Stage 1 summary statistics and the BDR dataset across the 500kb *APOE* region under investigation (Supplemental 1). Individual SNP AUC values within the BDR dataset identified 26 SNPs of interest, inclusive of the *APOE* isoform SNPs and spanning 92kb (92,040bp). Out of the 26 SNPs, 14 achieve AUC's of over 0.55, with 8 over 0.60; these mapped on to those SNPs which were found to be significantly associated with the AD phenotype in the BDR dataset (P<0.05). SNPs with AUC's ≥0.60 are all found within a 39kb region (39,220bp), in addition to 3 SNPs with AUCs ≥0.55 and 2 SNPs with AUCs less than 0.55. Both rs429358 (AUC=0.7025) and rs7412 (AUC=0.5304) are within this region, with their combined isoform predictability providing an AUC of 0.7082 ($P=6.89x10^{-15}$). Notably the region of significant SNPs and those that have higher AUCs are upstream to the *APOE* isoform SNPs with little indication for the involvement of SNPs downstream of the rs7412 SNP (Table 1).

Visualisation of the clump assignment of these SNPs at various levels of r^2, demonstrate two clear blocks of LD covering the *TOMM40* and *APOE* genes (Table 1). It was hypothesised that SNPs that were tagging each other would not alter the AUC statistic of the polygenic risk scores generated, as the additional effects sizes of the SNPs would alter the scores for both cases and controls within the same margin, therefore not altering the difference between them in the overall risk score. To test this, SNPs assigned to "clump one" were added into the polygenic risk score model as the level of r^2 decreased (Table 2). The SNPs incorporated into the PRS model up to an $r^2 \geq 0.6$ have similar effect sizes, leading to marginal changes in the AUC generated by these scores. As more SNPs with more variable effect sizes are added, the AUC also varies and increases, suggesting these SNPs may be independently contributing to the PRS.

A secondary block of SNPs (rs8106922-rs1160985-rs405509) consistently identified as belonging to the same clump (although the clump number changes across the r^2 parameters) was also observed and subjected to the same exploration to see if the same pattern of additional SNPs contributing to the model when lower levels of LD is utilised was seen. Again, this clump suggests that at lower parameters of r^2, additional SNPs are contributing to the model rather than tagging other SNPs (Table 3).

The discriminability of PRS consisting of the most significant SNP (by IGAP_Stage1 statistics) for each clump at various levels of linkage disequilibrium were compared to identify the most significant model, utilising only those SNPs whose individual AUC was greater than 0.55. This suggests that using an r^2 of ≥0.7 provides the most discriminatory model (Table 4) incorporating 10 SNPs within this region, although negligible differences in AUC were observed between the r^2 range of 0.5-0.7.

BP	45329214	45333834	45338895	45351746	45366779	45370941	45371168	45371328	45372354	45376284	45377467	45379060	45379791	45382034	45382675	45388130	45394336	45395266	45395909	45396219	45401666	45403412	45408836	45411941	45412079	45421254
SNP	rs10402271	rs4803760	rs34374273	rs1871047	rs519825	rs41290100	rs4803766	rs149529419	rs8104483	rs519113	rs2075642	rs387976	rs11667640	rs6859	rs41290120	rs34342646	rs71352238	rs157580	rs34404554	rs157582	rs8106922	rs1160985	rs405509	rs429358	rs7412	rs12721046
Beta in IGAP	0.271	-0.097	-0.192	-0.141	0.143	-0.570	-0.035	-0.190	-0.050	-0.011	-0.055	-0.120	-0.203	0.334	-0.608	1.104	1.045	-0.378	1.054	0.946	-0.381	-0.408	-0.303	1.350	-0.387	1.250
P value in IGAP	6.19E 60	1.26E 05	8.22E 12	2.20E 18	1.00E 18	2.82E 08	0.0293	0.04636	0.005597	0.583	0.004557	4.39E 12	1.12E 09	3.31E 96	1.45E 33	4.93e-440	2.49e-462	1.21E 101	6.41e-466	9.70e-434	2.04E 114	6.16E 125	4.35E 73	6.70e-536	1.23E 22	1.05e-421
Genes				NECTIN2														TOMM 40					APOE			
Clumps @																										APOC1
0.9	35	253	127	81	65	168	720	3242	847	-	770	121	153	22	58	1	4	19	1	2	17	16	32	9	73	12
0.8	26	211	104	67	33	137	580	2681	685	-	183	99	123	17	47	1	1	14	1	2	12	12	23	7	59	9
0.7	20	179	88	55	26	115	510	2311	594	-	154	83	90	14	39	1	1	12	1	2	10	10	10	5	49	7
0.6	17	151	54	46	22	100	452	2041	519	-	130	72	36	12	33	1	1	10	1	2	8	8	8	4	40	4
0.5	12	134	46	19	18	91	402	1805	117	-	117	32	31	11	28	1	1	9	1	1	7	7	7	2	35	5
0.4	10	104	53	33	15	75	15	1592	94	11175	94	50	26	9	23	1	1	6	1	1	6	6	6	1	29	4
0.3	9	87	45	27	8	64	34	1406	80	22	9	31	22	8	20	1	1	5	1	1	5	2	5	1	20	1
0.2	6	13	13	6	5	52	253	1190	5	44	5	17	44	5	15	1	1	3	1	1	3	3	3	1	15	1
0.1	3	8	20	2	5	29	721	969	3	8	2	2	9	1	8	1	1	2	1	1	2	2	2	1	8	1
BDR																										
AUC	0.5616	0.5268	0.5292	0.5362	0.5534	0.5276	0.5244	0.5174	0.5539	0.5005	0.5385	0.5058	0.5487	0.6108	0.4957	0.6847	0.6847	0.6178	0.6845	0.7098	0.5709	0.5843	0.5576	0.7025	0.5304	0.6904
Minor_Allele	G	T	A	G	C	T	A	A	G	G	A	C	T	A	A	A	C	G	C	T	G	T	G	C	T	A
MAF_Cases	0.376	0.243	0.074	0.361	0.399	0.021	0.445	0.007	0.252	0.277	0.170	0.346	0.053	0.524	0.038	0.323	0.323	0.284	0.319	0.409	0.295	0.312	0.427	0.411	0.049	0.340
MAF_Controls	0.296	0.204	0.104	0.409	0.329	0.046	0.476	0.024	0.314	0.268	0.207	0.360	0.104	0.372	0.037	0.104	0.104	0.451	0.101	0.146	0.399	0.436	0.509	0.168	0.082	0.114
Pvalue	0.011	0.164	0.114	0.141	0.031	0.025	0.361	0.019	0.037	0.778	0.146	0.654	0.003	5.15E-06	0.945	4.24E-14	4.24E-14	1.08E-07	4.30E-14	5.38E-17	0.001	9.66E-05	0.013	1.04E-14	0.037	2.07E-14
OR	1.437	1.255	0.695	0.818	1.352	0.443	0.885	0.283	0.736	1.043	0.783	0.940	0.488	1.859	1.025	4.126	4.126	0.482	4.184	4.028	0.629	0.586	0.718	3.467	0.578	4.031

Table1: Summary table of data collected from the IGAP_Stage1 summary statistics, clumping of the IGAP_Stage1 summary statistics at various levels of r^2, and association data and individual SNP AUC generated on the BDR dataset covering a 92kb region of interest. SNPs with high Area Under the Curve (AUC) measures align with association P values within the BDR dataset.

		Linkage Disequilbrium r^2 parameter					
	Individual	0.9	0.8-0.6	0.5	0.4	0.3-0.2	0.1
rs34404554	0.6845	0.6857	0.6857	0.7155	0.7313	0.7374	0.7429
rs34342646	0.6847	2.97×10^{-12}	3.26×10^{-12}	3.0×10^{-13}	5.4×10^{-15}	3.73×10^{-15}	3.32×10^{-15}
rs71352238	0.6847						
rs157582	0.7098						
rs429358	0.7025						
rs12721046	0.6904						
rs6859	0.6108						

Table 2: Summary table of AUC generated for the SNPs assigned as belonging to clump 1 in the IGAP_Stage 1 summary statistics at each level of r^2. Both SNPs rs34404554 and rs34342646 are in high LD and remain in the same clump throughout the r^2 parameters. Individual they provide similar AUC, which is not dissimilar when they are combined into the PRS. Addition of SNP rs71352238 throughout r^2 parameters of 0.8-0.6 does not see a significant change in the polygenic risk score statistics. However additional SNPs added at lower levels of linkage disequilibrium appear to improve both the significance of the PRS and discriminability.

		Linkage Disequilbrium r^2 parameter				
	Individual	0.9	0.8	0.7-0.5	0.4 - 0.2	0.1
rs1160985	0.5843	0.00031	0.5809	0.5737	0.6581	0.6686
rs8106922	0.5709		0.00064	0.0013	2.04×10^{-8}	2.6×10^{-9}
rs405509	0.5576					
rs157580	0.6178					
rs1871047	0.5362					
rs387976	0.5058					
rs2075642	0.5385					

Table 3: Secondary block of SNPs in strong LD across markers rs8106922-rs1160985-rs405509 displayed no improvement of AUC when SNPs were in strong LD, however at lower levels of r^2 (<0.4) additional SNP contributed to the model.

The BDR target dataset was divided into two groups based on whether the individual was *APOE* ε4 allele positive or not. This resulted in a dataset of 247 cases and 53 controls in the *APOE* ε4 allele positive group (ε4 pos): and 108 cases and 111 controls in the *APOE* ε4 allele negative group (ε4 neg). The 10 SNP PRS model was applied to these sub-groups and as expected the AUC in the *APOE* ε4 allele positive was not dissimilar to that of the full dataset (AUC 0.7368 v 0.7477 in the full dataset) whilst the score for those with no *APOE* ε4 alleles display minimal discriminability (0.5284), confirming the presence of the *APOE* ε4 allele is a strong contributory factor in predicting AD (Table 5, header row).

To test if each SNP within this 10 SNP model was contributing to the 0.7477 AUC, a drop-out analysis was carried out, removing a single SNP from the model, and observing if a drop of greater than 0.005 was observed in the AUC. The removal of SNPs, rs34404554 and rs1457582 that reside within "clump 1" of the IGAP_S1 summary statistics across all r^2 parameters does not result in a large drop in the AUC when analysed in the entire dataset or in the *APOE* ε4 allele positive/negative datasets. Likewise, SNPs, rs1160985, rs104022771, rs519825 and rs8104483 are not making strong contributions to the discriminatory value of the 10 SNP model. However large drops in AUC are observed when the *APOE* isoform SNP rs429358 is removed from the model, supporting its role as a key contributory factor. In addition, removal of rs157580, consistently reduced the predictability of the model across full dataset and in the *APOE* ε4 allele positive/negative sub-groups, suggesting this SNP may also have a contributory role independent of the rs429358 isoform SNP (Table 5). The roles of SNPs rs12721046 and rs6859 are less clear, although lower AUCs are observed when these SNPs are removed from the entire dataset and in the *APOE* ε4 allele positive it does not make the >0.005 cut-off, however in the *APOE* ε4 allele negative dataset, the removal of these SNPs from the model indicate key contributory roles, which may only be observed in the absence of the rs429358 *APOE* ε4 isoform SNP.

To confirm the findings of the drop-out analysis, a "drop-in" analysis was performed adding in the key contributory SNPs identified to observe the impact on the AUC (Table 6). The resultant 4 SNP model had a higher AUC than the original 10 SNP model, indicating that the removal of SNPs that do not contribute to the model, reduces noise, and improves the discriminatory accuracy. However, this was only observed in the entire dataset and the *APOE* ε4 allele negative dataset. In the *APOE* ε4 allele positive sub-group the original model fairs better.

Individual AUC ≥0.55			
r2	# SNP	P value	AUC
All SNPs	14	4.63 x 10-15	0.731
0.9	13	1.95 x 10-15	0.7332
0.8	11	3.79 x 10-16	0.7393
0.7	**10**	**1.53 x 10-16**	**0.7477**
0.6	9	1.55 x 10-16	0.7456
0.5	9	2.95 x 10-16	0.7466
0.4	7	4.96 x 10-16	0.7393
0.3	5	1.94 x 10-13	0.7115
0.2	4	2.33 x 10-15	0.7271
0.1	4	1.03 x 10-14	0.7162

Table 4: Table showing the results of polygenic risk score models consisting of the most significant SNP within each clump in the region of interest at each r^2 parameter. For this only SNPs with individual AUCs of greater than 0.55 were included. This suggests that an r^2 parameters of ≥0.7 provides the most discriminatory model, though there is minimal different in AUC between r^2 measures of 0.5-0.7.

Dropped SNP	r^2 parameter SNP enters "Clump 1"	All		ε4 pos		ε4 neg	
		P value	AUC	P value	AUC	P value	AUC
		1.53x10-16	0.7477	1.23x10-06	0.7368	0.18	0.5284
rs429358	**0.4**	1.76x10-15	0.7419	6.38x10-07	0.7315	0.18	0.5284
rs34404554	**0.9**	3.88x10-17	0.7507	7.21x10-07	0.7416	0.17	0.5287
rs157582	**0.5**	2.9x10-16	0.7468	5.00x10-06	0.7342	0.12	0.5248
rs12721046	**0.3**	1.55x10-16	0.7456	7.91x10-07	0.7494	0.39	0.5109
rs1160985		1.34x10-16	0.7527	1.79x10-06	0.7322	0.11	0.5586
rs157580		4.02x10-16	0.7382	2.19x10-06	0.7291	0.3	0.5111
rs6859	**0.1**	1.58x10-16	0.7453	1.58x10-06	0.7357	0.21	0.476
rs10402271		2.26x10-16	0.7459	2.00x10-06	0.7296	0.17	0.5312
rs519825		1.18x10-16	0.7491	1.17x10-06	0.7383	0.16	0.5342
rs8104483		1.51x10-16	0.7472	1.22x10-06	0.7383	0.19	0.5243

Table 5: Summary of "drop-out" analysis to determine the level of contribution each SNP is having to the polygenic risk score model. Grey shading indicates drops in AUC of ≥0.005 when SNP is removed from the model. This analysis that SNPs rs429358 and rs157580 are important contributory SNPs in the polygenic risk score model for AD. In the absence of the *APOE* ε4 allele, two further SNPs, rs12721046 and rs6859, which display a low level of linkage disequilibrium with the *APOE* isoform SNP rs429358 (r^2<0.4), are also important contributing SNPs in the discriminatory model.

	All		ε4 pos		ε4 neg	
	P value	AUC	P value	AUC	P value	AUC
10 SNP r^2≥ model	1.53x10-16	0.7477	1.23x10-06	0.7368	0.18	0.5284
rs429358	7.39 x 10-15	0.7025	0.0188	0.5722	-	-
+ rs157580	8.40 x 10-16	0.7283	0.0021	0.6406	0.1376	0.5504
+ rs6859	4.59 x 10-16	0.7398	0.0003	0.6843	0.1399	0.5595
+ rs12721046	1.00 x 10-16	0.7532	2.76 x 10-5	0.7114	0.039	0.5775

Table 6: Drop-in analysis, adding in contributory SNPs identified singly to obtain a 4-SNP model that provides a greater discriminatory accuracy than the 10 SNP model in the entire dataset the *APOE* ε4 allele negative sub-group. However, no improvement was observed in the *APOE* ε4 allele positive sub-group.

For completeness, rs7412 was added into this 4-SNP model to ensure capture of the full *APOE* isoform effect. Inclusion of this SNP only marginally changed the discriminatory accuracy of the 4-SNP model with AUCs of 0.7525, 0.7134, and 0.5746 obtained for the full cohort and *APOE* ε4 allele positive and negative subgroups respectively.

Discussion

The presence of the *APOE* ε4 allele is one of the strongest risk factors associated with the onset of AD[1,2], due to this many PRS investigations on AD, remove a substantial region of genotype data surrounding this locus due to the assumption that additional association signals in this region are due to the SNPs being in LD with the isoform SNPs rs429358 and rs7412. This study has investigated this region of chr19, utilising PRS analysis and clumping algorithms to determine if by removing this region of the genome predictive SNPs for AD are being lost. The data presented here suggest that the "sphere of influence" of the *APOE* isoform SNPs covers a region of around 92kb; SNPs in LD at r^2<0.4 with rs429358 potentially contribute to the PRS predictability for AD, and that there are additional independent SNPs in this region that demonstrate independent contributory effects in an *APOE* ε4 negative sample.

AUC statistics obtained for the *APOE* isoform SNPs in the BDR sample present here provide an overall model accuracy of 0.7082, with the increase in score significantly correlated to disease outcome ($P=6.89x10^{-15}$). This is comparable to PRS outcomes observed in other studies[8,27–29], confirming the BDR cohort is showing the same genetic architecture of larger cohorts.

LD is where there is a non-random association of alleles between loci, suggesting they are co-inherited at a frequency that is higher than chance. The r^2 parameter indicates the level at which 2 alleles are correlated, and therefore when one is known provides an approximate prediction of the allele at the second loci, with r^2=1 indicating the two alleles are in perfect correlation, with a certainty of allelic prediction, and 0 suggesting no correlation, and therefore random chance of allele prediction. Consequently, the higher the r^2 between two SNPs the more correlated they are and the more accurate the allele prediction will be. Using this assumption many PRS studies opt to clump the SNP dataset being used so that only a single (most significant) SNP is used from a haplotype where the correlation (or r^2) between them is ≥0.1, although the genetic distance in which the SNP lie in proximity to each other to assess varies from 250kb (default PLINK parameter) to 1000kb windows[7,8,10,11,20,30,31]. Although this almost certainly ensures independence of the SNPs being entered into the PRS model, it could also lead to many key independent SNPs being omitted from the analysis. Traditionally "tag" SNPs were genetic variants that were genotyped as proxies for additional SNPs in high LD around them, reducing redundancy in genotyping efforts for GWAS studies, investigations into the selection of these "tag" SNPs to capture the maximum variation of the genome suggest using SNPs with r^2 ≥0.5[32,33].

In this analysis we clumped the base dataset and labelled each SNP with the clump identifier it belongs to at each r^2 parameter. This demonstrated that the SNPs within the 500kb surrounding the *APOE* are not in high LD, with only 2 SNPs sharing at single clump at the highest r^2 parameter of ≥0.9. As the r^2 metric decreases, more SNPs are assigned to the same clump and by r^2≥0.1, the majority of the 26 SNPs are tagged by 3 haplotype blocks. When exploring the 2 main haplotype blocks with the immediate region upstream to the isoform SNPs, including SNPs with higher levels of LD (>0.5) into the PRS models, does not greatly alter the AUCs observed (Tables 2 and 3). It is only when SNPs with lower levels of r^2 are added into the model is there an increase in predictability. This suggests that a) when SNPs in high LD are included together in PRS models it does not artificially increase the AUC and b) the current practise of clumping SNPs at an r^2 ≥0.1 may be forcibly removing SNPs that could be contributing to the phenotype. Interesting, the LD level at which additional SNPs are informative to these models is around an r^2 of 0.5; which is the same level suggested by those involved in the HapMap and Tag SNP selection algorithms suggest is required to sufficiently capture the variation in the surrounding genome[32,33].

This study has identified 3 potentially contributory SNPs to the PRS (in addition to rs429358) that are likely

to be omitted from analyses. The SNP rs157580, becomes part of the secondary haplotype block (Table 3) when the clumping of SNP occurs with r^2 less than 0.5, and therefore was included in as an independent SNP in the best observed model including the most significant SNPs in each clump at an r^2 of ≥0.7. Drop-out analysis consistently suggested that the SNP was making contributory effects to the PRS model in the whole dataset and in *APOE* ε4 sub-groups. The minor G-allele of this SNP is less frequent in cases than in controls and therefore has a beta effect size of -0.378 in the IGAP_S1 summary statistics ($P=1.21 \times 10^{-101}$). The SNP resides within the *TOMM40* gene which has been hypothesized as having an association with AD via an intron 6 poly-T variant[17,34–36], although not consistently observed across studies[37,38], and has been attributed to the level of LD observed between this variant and the *APOE* isoform[39,40]. Rs157580 is located further upstream within intron 1 and has been observed to be association with AD in multiple studies[14,15], potentially altering intron excision rates[16].

Further to this two SNPs that also clumped with one of the two major haplotypes within this region also appeared to make independent contributions to the PRS. SNPs rs6859 and rs12721046 reside on opposite ends of the region of interest spanning the *APOE* isoform SNPs, with both being tagged by the rs429358 SNP at low levels of LD (r^2 of 0.1 and 0.3 respectively). The rs12721046 resides in the downstream gene to *APOE*, Apolipoprotein C1 (*APOC1*), and has also been shown to be associated with the AD phenotype[19,41,42]. A recent study[43] suggests that a haplotype across the *APOE* locus consisting of rs2075650 (*TOMM40*) - *APOE* ε4 – rs12721046 (*APOC1*), has a stronger association with AD than the ε4 allele alone. This would support the suggestion from this investigation that *TOMM40* contributes to the AD phenotype, however the SNP Kulminski and colleagues[43] identified, although is not in the BDR analysis and so therefore is absent from our analysis, on exploration of the clump tagged IGAP_S1 dataset, it was found that rs2075650 resided in the linkage disequilibrium block "clump1" at all measures of r^2 (0.1-0.9) and suggests it may be part of the association block of rs429358.

The SNP rs6859 which resides in the *NECTIN2/PVRL2* gene upstream to the *TOMM40-APOE-APOC1* LD block, has also been identified in several AD investigations[15,18,19,44–48]. The data present here would suggest that both rs6859 and rs12721046 are only contributing to the predictability of AD in the absence of the ε4 allele. The increase in AUC in the whole datasets and the ε4 negative sub-group, but negligible change in the ε4 positive subgroup when they are removed from the PRS 10 model demonstrates this and is supported by the improved predictability observed in the 4 SNP haplotype model in those same datasets, possibly suggesting that the large effect size of the ε4-allele overshadows their effects. This is support by the observations by Zhou and colleagues exploration of the *APOE* region, observing both the rs6859 and rs12721046 were associated in samples homozygous for ε3 allele[19]. This along with the data presented here suggests that some key variants in the *APOE* locus may be independently contributing to the AD phenotype and consist of a disease risk haplotype when in combination with the ε4 allele. It is perhaps because of the high frequency of the ε4 allele found in AD cohorts that these variants have been overlooked and be obscuring association and PRS analyses, as on an ε4 positive background it would seem these SNPs have little effect. Additional studies on this region, and a deeper look into the haplotypes is warranted, especially in AD cases that do not carry the ε4 allele, and across ethnic groups.

Acknowledgements

We would like to gratefully acknowledge all donors and their families for the samples provided for the BDR cohort and additional datasets who genetic data was also utilised here.

Tissue samples from the BDR cohort were obtained from the Southwest Dementia Brain Bank, London Neurodegenerative Diseases Brain Bank, Manchester Brain Bank, Newcastle Brain Tissue Resource and Oxford Brain Bank, and we thank our colleagues of the BDR Network, in particular the neuropathologists at each centre and BDR Brain Bank staff for the collection and classification of the samples. The BDR is jointly funded by Alzheimer's Research UK and the Alzheimer's Society in association with the Medical Research Council. Brains for Dementia Research has ethics approval from London – City and East NRES committee 08/H0704/128+5 and has deemed all approved requests for tissue to have been approved by the committee.

The genotyping of the Brains for Dementia cohort is supported by funding provided by an ARUK project grant, entitled 'Enabling high-throughput genomic approaches in Alzheimer's disease' and an ARUK extension grant entitled 'NeuroChip analysis of the entire Brains for Dementia Research (BDR) resource of 2000 samples', awarded to KJB.

Conflict of Interest Statement

The author declares there is no conflict of interest.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

References

1. Pericak-Vance MA, *et al.* Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *Am J Hum Genet* **48**, 1034–1050 (1991).

2. Corder EH, *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science (80-.).* **261**, 921–923 (1993).

3. Kukull WA, *et al.* Apolipoprotein E in Alzheimer's disease risk and case detection: A case-control study. *J. Clin. Epidemiol.* **49**, 1143–1148 (1996).

4. Farrer LA. *et al.* Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* **278**, 1349–1356 (1997).

5. Brookes KJ & Morgan K. Genetics of Alzheimer's Disease. *eLS.* (2017) doi: 10.1002/9780470015902.a0020228.pub2.

6. Baker E & Escott-Price V. Polygenic Risk Scores in Alzheimer's Disease: Current Applications and Future Directions. *Front. Digit. Heal.* **2**, 14 (2020).

7. Fulton-Howard B, *et al.* Greater effect of polygenic risk score for Alzheimer's disease among younger cases who are apolipoprotein E-ε4 carriers. *Neurobiol. Aging* **99**, 101. e1-101. e9 (2021).

8. Leonenko G, *et al.* Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nat. Commun.* **12**, 4506 (2021).

9. Escott-Price V, Shoai M, Pither R, et al. Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. *Neurobiol Aging* **49**, 214 e7-214 e11 (2017).

10. Najar J, *et al.* Polygenic risk scores for Alzheimer's disease are related to dementia risk in APOE ε4 negatives. **13**, e12142 (2021).

11. Stocker H, *et al.* Prediction of clinical diagnosis of Alzheimer's disease, vascular, mixed, and all-cause dementia by a polygenic risk score and APOE status in a community-based cohort prospectively followed over 17 years. *Mol. Psychiatry 2020 2610* **26**, 5812–5822 (2020).

12. Logue MW, *et al.* Use of an Alzheimer's disease polygenic risk score to identify mild cognitive impairment in adults in their 50s. *Mol Psychiatry* (2018) doi: 10.1038/s41380-018-0030-8.

13. Blom ES, *et al.* Does APOE explain the linkage of Alzheimer's disease to chromosome 19q13? *Am J Med Genet B Neuropsychiatr Genet* **147B**, 778–783 (2008).

14. Potkin SG, *et al.* Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS One* **4**, e6501 (2009).

15. Jia L, *et al.* Prediction of Alzheimer's disease using multi-variants from a Chinese genome-wide association study. *Brain* **144**, 924–937 (2021).

16. Belloy ME, Eger SJ, Guen Y Le, et al. Two APOE splice sQTLs reduce Alzheimer's disease risk in APOE 4/4 carriers. *Alzheimer's Dement.* **16**, e043539 (2020).

17. Roses AD, *et al.* A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. *Pharmacogenomics J.* **10**, 375 (2010).

18. Logue MW, *et al.* A comprehensive genetic association study of Alzheimer disease in African Americans. *Arch. Neurol.* **68**, 1569–1579 (2011).

19. Zhou X, *et al.* Non-coding variability at the APOE locus contributes to the Alzheimer's risk. *Nat. Commun.* **10**, 1–16 (2019).

20. Chaudhury S, *et al.* Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment. *Transl. Psychiatry* **9**, 154 (2019).

21. Lambert J-C, *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).

22. Young J, *et al.* Genome-wide association findings from the brains for dementia research cohort. **107**, 159–167 (2021).

23. 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).

24. R Core team. R: A language and environment for statistical computing. (2021).

25. Robin X, *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

26. Makiyama K. magicfor: Magic Functions to Obtain Results from for Loops. (2016).

27. Escott-Price V, *et al.* Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain* **138**, 3673–3684 (2015).

28. Chaudhury S, *et al.* Polygenic risk score in postmortem diagnosed sporadic early-onset Alzheimer's disease. *Neurobiol. Aging* **62**, 244. e1-244.e8 (2018).

29. Daunt P, *et al.* Polygenic Risk Scoring is an Effective Approach to Predict Those Individuals Most Likely to Decline Cognitively Due to Alzheimer's Disease. *J. Prev. Alzheimer's Dis.* **8**, 78–83 (2021).

30. Logue MW, *et al.* Use of an Alzheimer's disease polygenic risk score to identify mild cognitive impairment in adults in their 50s. *Mol. Psychiatry 2018 243* **24**, 421–430 (2018).

31. Escott-Price V, *et al.* Genetic analysis suggests high misassignment rates in clinical Alzheimer's cases and controls. *Neurobiol. Aging* **77**, 178–182 (2019).

32. De Bakker PIW, *et al.* Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.* **38**, 1298–1303 (2006).

33. Carlson CS, *et al.* Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium. *Am. J. Hum. Genet.* **74**, (2004).

34. Johnson SC, *et al.* The Effect of TOMM40 Poly-T length on Gray Matter Volume and Cognition in Middle-Aged Persons with APOE ε3/ε3 Genotype. *Alzheimers. Dement.* **7**, 456 (2011).

35. Roses AD, An inherited variable poly-T repeat genotype in TOMM40 in Alzheimer disease. *Arch. Neurol.* **67**, 536–541 (2010).

36. Maruszak A, *et al.* TOMM40 rs10524523 polymorphism's role in late-onset Alzheimer's disease and in longevity. *J. Alzheimers. Dis.* **28**, 309–322 (2012).

37. Cruchaga C, *et al.* Association and expression analyses with single-nucleotide polymorphisms in TOMM40 in Alzheimer disease. *Arch. Neurol.* **68**, 1013–1019 (2011).

38. Jun G, *et al.* Comprehensive search for Alzheimer disease susceptibility loci in the APOE region. *Arch. Neurol.* **69**, 1270–1279 (2012).

39. Linnertz C, *et al.* Characterization of the Poly-T Variant in the TOMM40 Gene in Diverse Populations. *PLoS One* **7**, e30994 (2012).

40. Guerreiro RJ & Hardy J. TOMM40 association with Alzheimer disease: tales of APOE and linkage disequilibrium. *Arch. Neurol.* **69**, 1243–1244 (2012).

41. Gao L, Cui Z, Shen L, et al. Shared Genetic Etiology between Type 2 Diabetes and Alzheimer's Disease Identified by Bioinformatics Analysis. *J. Alzheimer's Dis.* **50**, 13–17 (2016).

42. Zhou X, *et al.* Identification of genetic risk factors in the Chinese population implicates a role of immune system in Alzheimer's disease pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1697–1706 (2018).

43. Kulminski AM, Philipp I, Shu L, et al . Definitive roles of TOMM40-APOE-APOC1 variants in the Alzheimer's risk. *Neurobiol. Aging* **110**, 122–131 (2022).

44. Xiao Q, *et al.* The Relationship Between Low-Density Lipoprotein Cholesterol and Progression of Mild Cognitive Impairment: The Influence of rs6859 in PVRL2. *Front. Genet.* **13**, 194 (2022).

45. Xiao Q, *et al.* Risk prediction for sporadic Alzheimer's disease using genetic risk score in the Han Chinese population. *Oncotarget* **6**, 36955–36964 (2015).

46. Zhou X, *et al.* Genetic and polygenic risk score analysis for Alzheimer's disease in the Chinese population. **12**, e12074 (2020).

47. Rao S, *et al.* An APOE-independent cis-eSNP on chromosome 19q13.32 influences tau levels and late-onset Alzheimer's disease risk. *Neurobiol. Aging* **66**, 178. e1-178.e8 (2018).

48. Jun G, *et al.* Comprehensive Search for Alzheimer Disease Susceptibility Loci in the APOE Region. *Arch. Neurol.* **69**, (2012).