Analysis of whole genome sequenced cases and controls shows that the association of variants in TOMM40, BCAM, NECTIN2 and APOC1 with late onset Alzheimer's disease is driven by linkage disequilibrium with APOE ε2/ε3/ε4 alleles.

David Curtis and the Alzheimer’s Disease Neuroimaging Initiative

UCL Genetics Institute, UCL, Darwin Building, Gower Street, London WC1E 6BT.
Centre for Psychiatry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ.

Correspondence:
David Curtis d.curtis@ucl.ac.uk
00 44 7973 906 143

Running head: Role of APOE in Alzheimer’s disease

Abstract

Variants in APOE are associated with risk of late onset Alzheimer's disease (LOAD) but the magnitude of the effect has been reported to vary across ancestries. Also, other variants in the region have been reported to show association though it has been unclear whether this was secondary to their linkage disequilibrium with the APOE variants rs429358 and rs7412. Previous analyses of exome-sequenced samples have identified other genes in which rare variants impact risk of disease. In this study 2000 whole genome sequenced cases and controls with different ancestries were subjected to gene-based weighted burden analysis to identify risk genes. Additionally, individual variants in the APOE region were tested for association with LOAD. When using the APOE variants as covariates no individual genes showed statistically significant evidence for association after Bonferroni correction for multiple testing, which may well be a consequence of the modest sample size. Likewise, for those variants initially showing evidence of association with LOAD incorporating the APOE variants as covariates dramatically reduced the strength of association. These results demonstrate that the differential association of APOE across ancestries does not appear to be driven by another variant in the region. It seems likely that no other genes in the region have a direct effect on LOAD risk.

Keywords

LOAD; BCAM; NECTIN2; TOMM40; APOC1; APOE.

Introduction
The first report of association between APOE alleles and late onset Alzheimer’s disease (LOAD) noted that there were three isoforms of the ApoE protein, referred to as ApoE-E2, ApoE-E3 and ApoE-E4, and that restriction-based isotyping showed that the frequency of the APOE ε4 allele in 30 unrelated late onset familial cases was significantly different from that in 91 controls (0.50 v. 0.16, p=0.014) (Strittmatter et al., 1993). A table in the same paper showed the allele frequency of the APOE ε2 allele to be 0.04 in cases versus 0.08 in controls although this was not commented on at the time but a protective effect for the APOE ε2 allele was reported the following year (Corder et al., 1994). Apo-E4 contains an arginine at residue 112 whereas E3 has a cysteine and the variant encoding this substitution is now named rs429358:T>C. Apo-E2 contains a cysteine at residue 158 while E3 has an arginine and this is coded for by rs7412:C>T. The two variants are in strong LD with each other and it is thought that the ε3 allele derived from the ancestral ε4 allele and that the ε2 allele subsequently derived from the ε3 allele, with the ε4 allele possibly being under negative selection owing to its association with a less favourable lipid profile and increased risk of cardiovascular disease (Fullerton et al., 2000).

The association of these APOE variants with LOAD was confirmed and functional studies in mice have shown that the human APOE isoforms are associated with differences in amyloid plaque deposition (E4 > E3 > E2) and differences in clearance of amyloid-β (Aβ) peptide (E2 > E3 > E4) (Bales et al., 2009; Castellano et al., 2011). A very rare allele of APOE is referred to as the APOE3 Christchurch (R136S) variant, GRCh38:g.19-44908756C>A (Wardell et al., 1987). There is a recent case report of an elderly subject with the pathogenic PSEN1 E280A variant, an autosomal dominant cause of presenile Alzheimer’s disease, who was also homozygous for the APOE3 Christchurch variant and who had high brain amyloid levels with only mild cognitive impairment (Arboleda-Velasquez et al., 2019). Investigations revealed she had high amyloid-β plaque burden but limited neurodegeneration or tau accumulation. In vitro experiments showed that the APOE3ch isoform triggered less Aβ42 aggregation than APOE3, suggesting a possible mechanism underlying a protective effect. They also showed that APOE3ch has the lowest heparin binding ability of all APOE isoforms, whereas APOE4 had been shown to have a higher heparin affinity than APOE3 (Yamauchi et al., 2008).

Although the association of APOE variants with LOAD was repeatedly replicated, differences related to ancestry were soon reported. For purposes of clarity, the original nomenclature is retained here even though it might differ from the language we would use to describe ancestry effects in today’s literature. An early meta-analysis reported that the APOE ε4 AD association was weaker among African Americans and Hispanics than for Caucasians (Farrer et al., 1997) and a recent study confirmed that rs429358 was more strongly associated with LOAD in whites than in Hispanics, with the effect in African-Americans being intermediate (Kulminski et al., 2019). This study analysed variants from the region including BCAM, NECTIN2, TOMM40, APOE, and APOC1 and observed that rs429358 and rs7412 were in LD with nearby SNPs and that the patterns of LD varied between ethnicities. If nearby SNPs independently influenced LOAD risk then such differences could theoretically account for
different strengths of association between rs429358, the APOE ε4 variant, and LOAD in different ancestries. For example, there have been inconsistent claims that rs449647 and rs405509 are independently associated with LOAD, with one study reporting that the haplotype of rs405509-T with rs429358-C increased risk whereas the rs405509-G with rs429358-C did not, although this effect was restricted to subjects aged over 75 (Lescai et al., 2011). A study in 525 whole genome sequenced Chinese subjects and large samples of Chinese and non-Asian subjects identified haplotypes in the region exerting independent effects on risk, apparently exerted via modulating gene expression, but this study did not explicitly address the question of whether these haplotypes might explain the differential effects of APOE ε4 in different ancestries (Zhou et al., 2019). The haplotypes reported were made up of multiple non-coding variants, none of which individually had a substantial effect.

Two additional recent studies have demonstrated that local ancestry around APOE moderates the risk associated with the APOE alleles, again suggesting that other nearby variants might have an effect (Blue et al., 2019; Rajabli et al., 2018).

One way to explain the observation that rs429358-C is more strongly associated with LOAD risk in Caucasians than Hispanics would be to postulate another locus influencing risk in linkage disequilibrium (LD) with it. LD describes the situation where there are two polymorphic loci which are close together on the same chromosome for which particular pairs of alleles occur together in the same haplotype more often than would be expected by random segregation. Envisioning this scenario, the high risk allele at the second locus might tend to occur in the same haplotype as rs429358-C in Caucasians and in the same haplotype as rs429358-T in Hispanics. Thus it would amplify the effect of rs429358-C in Caucasians and counteract it in Hispanics. If the LD between the loci was strong within each group then in a homogeneous sample of either Caucasians or Hispanics it might not be possible to detect an independent effect of the second variant. However if one analysed a sample consisting of a mixture of Caucasian and Hispanic cases and controls then in a multivariate analysis the separate contributions of both loci should become apparent. The present report describes such analyses in a mixed sample. The purpose of these analyses was firstly to determine whether it was possible to identify genes apart from APOE affecting LOAD risk and secondly to determine whether variants close to APOE exerted independent effects on LOAD risk, especially in a way that could explain differences in APOE associations between different ancestries.

Methods

Data used in the preparation of this article was obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early
Alzheimer’s disease. 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 (Hurko et al., 2012). 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.

Phenotype information and variant calls based on whole genome sequencing using the GRCh38 assembly for case-control subjects of the extension phase of the Alzheimer's Disease Sequencing Project (ADSP) were downloaded from the NIAGADS website (https://www.niagads.org/adsp/content/study-design). Ethical approval and informed consent had been obtained by the researchers who generated this dataset. As described previously, participants were at least 60 years old and although different diagnostic procedures were used by different contributing studies all cases met NINCDS-ADRDA criteria for possible, probable or definite AD based on clinical assessment, or had presence of AD (moderate or high likelihood) upon neuropathology examination (Beecham et al., 2017). The subjects are described as being of Non-Hispanic White (NHW), Caribbean Hispanic (CH), and African American (AA) descent. Whole genome sequencing was carried out using standard methods as described previously (Naj et al., 2019; Vardarajan et al., 2018) and on the ADSP website: https://www.niagads.org/adsp/content/sequencing-pipelines. The PrevAD (prevalent Alzheimer's disease) field was used to define phenotype and subjects who had been included in the earlier exome-sequenced datasets were excluded, yielding 985 cases and 2,101 controls.

To obtain population principal components, version 1.90beta of plink (https://www.cog-genomics.org/plink2) was run on the chromosome 22 genotypes with the options --maf 0.1 -pca header tabs --make-rel (Chang et al., 2015; Purcell et al., 2007, 2009). This is a standard approach whereby the genotypes of a large number of loci are entered into a principal components analysis. The main principal components obtained are taken to reflect any geographical and ancestry-related sub-stratification within the samples and incorporating them into association analyses as covariates can mitigate problems caused by inadequate matching of cases and controls.

The analyses were carried out in two stages. Initially a weighted burden analysis of each gene was carried out using GENEVARASSOC and SCOREASSOC in the same way as had previously been applied to the ADSP exome-sequenced case-control samples (Curtis et al., 2019). For each gene, variants in the exomes and splice regions were extracted and annotated using VEP, PolyPhen and SIFT (Adzhubei et al., 2013; Kumar et al., 2009; McLaren et al., 2016). These programs assess the likely impact of DNA changes on the protein coded for by the gene and in particular whether the changes produced are likely to disrupt the normal functioning of the protein. SCOREASSOC was then used to carry out a weighted burden analysis to test whether, in a particular gene, variants which were rarer and/or predicted to have more severe functional effects occurred more commonly in cases than
Controls. Variants were weighted according to their functional annotation using the default weights provided with the GENEVARASSOC program, which was used to generate input files for weighted burden analysis by SCOREASSOC (Curtis, 2016, 2012). For example, a weight of 5 was assigned for a synonymous variant, 10 for a non-synonymous variant and 20 for a stop gained variant. Additionally, 10 was added to the weight if the PolyPhen annotation was possibly or probably damaging and also if the SIFT annotation was deleterious, meaning that a non-synonymous variant annotated as both damaging and deleterious would be assigned an overall weight of 30. The full set of weights is shown in Supplementary Table 1, copied from the previous report (Curtis et al., 2018). Variants were excluded if they did not have a PASS in the information field or if there were more than 10% of genotypes missing in either cases or controls or if the heterozygote count was smaller than both homozygote counts in both cohorts. For each subject a gene-wise weighted burden score was derived as the sum of the variant-wise weights, each multiplied by the number of alleles of the variant which the given subject possessed. If a subject was not genotyped for a variant then they were assigned the subject-wise average score for that variant.

As described previously, ridge regression analysis with lambda=1 was used to test whether the gene-wise score was associated with caseness (Curtis et al., 2018). To do this, SCOREASSOC first calculates the likelihood for the phenotypes as predicted by the first 20 principal components and then calculates the likelihood using a model which additionally incorporates the gene-wise scores. It then carries out a likelihood ratio test assuming that twice the natural log of the likelihood ratio follows a chi-squared distribution with one degree of freedom to produce a p value. Results are expressed as signed log p (SLP) which is the logarithm base 10 of this p value and is positive if the scores are higher in cases and negative if they are higher in controls. This analysis was carried out for every autosomal gene listed in the RefSeq GRCh38. All the analyses were then repeated using rs429358 and rs7412 genotypes as covariates as well as the principal components.

Data handling and generic statistical analyses were carried out using R (R Core Team, 2014).

In the initial weighted burden analyses without including rs429358 and rs7412 as covariates only two genes showed evidence for association which was statistically significant after Bonferroni correction for the number of genes tested, APOE and the neighbouring gene TOMM40. Therefore a second set of analyses was carried out in order to explore whether the result obtained for TOMM40 was likely to be due to direct functional effects of variants in this gene or whether it was more likely a consequence of LD relationships with APOE variants. Because previous reports have also claimed that variants in the neighbouring genes APOC1, BCAN and NECTIN2 might influence LOAD risk, all variants in the region including these five genes plus a margin of 10 kb at either end were extracted, between genomic coordinates 44799059 and 44929344. For each of the 718 variants with a minor allele frequency of at least 0.01 in cases and/or controls in the combined sample an individual ridge regression analysis was carried out including population principal components and
testing the effect of the variant with and without also including as covariates the genotypes for rs429358 and rs7412. In order to assess ancestry effects, analyses were carried out without including population principal components and raw odds ratio (OR) for association with LOAD was obtained for each variant. All variants with OR>2 or OR<0.6 were selected for further analysis. For these variants showing association with LOAD, the count of the number of alternate alleles was obtained and this allele count was then treated as a quantitative variable in joint ridge regression analyses with PrevAD as the outcome. In an attempt to identify those variants which might be exerting an independent effect on risk, a subset of 14 variants was selected using an arbitrary threshold of $|\beta/SE(\beta)| > 1.3$, with beta being the coefficient for the variant in the fitted multivariate model.

In order to explore whether any of these variants was exerting a direct effect on LOAD risk, likelihood ratio tests were carried out for each associated variant by comparing the likelihoods of models which did or did not include the variant in question, and which included the following as covariates:

1. None.

2. Each of the 14 variants identified above as possibly having an independent effect.

3. Rs429358 and rs7412 together.

These analyses were then repeated for each of the three ancestry groups separately. The results of these analyses are expressed as minus log p (MLP), which is the minus logarithm base 10 of the p value for the likelihood ratio test.

In order to explore the possible independent contributions of individual variants further, the subjects were subdivided according to the rs429358 genotype. Based on allele counts in cases and controls an odds ratio and Pearson chi-squared statistic were calculated for each variant in each group. Again, this process was then repeated with the subjects additionally divided by ancestry groups.

**Results**

Table 1 presents the age and sex distribution of the samples used. The weighted burden tests evaluated 1,662,028 variants in 28,862 autosomal genes. *TOMM40* produced an SLP of 20.43 and *APOE* produced an SLP of 13.1. No other gene produced results which would be considered statistically significant after correction for the number of genes tested but it may be worth noting that *APOC1* had the fourth highest SLP, of 4.2, and that *TREM2* had the eighth highest SLP, of 3.5. When these analyses were repeated using rs429358 and rs7412 as covariates no gene produced significant results. The SLP for *APOE* was 0.93 and for *TOMM40* was 0.86, suggesting no evidence for an independent effect of any other variants. The SLP of *APOC1* reduced to 1.94 while the SLP for *TREM2* fell only slightly to 3.3, ninth highest of all genes. No subject carried an *APOE3* Christchurch variant.
There were 718 genotyped common variants in the region between 10kb proximal of BCAM and 10kb distal of APOC1. In the analysis including principal components, 91 produced an MLP greater than 4.15, the critical value to be considered significant at p<0.05 after correction for multiple testing, when the APOE alleles were not included as covariates. For all of these the MLP reduced, sometimes dramatically, when APOE alleles were included in the model, indicating that initial result had been driven via LD. The highest MLP obtained after inclusion of the APOE alleles was 3.72 for rs59007384 and Table 2 shows the results for all the variants which produce an MLP of greater than 2.5 in the analyses including APOE. Important, none of the 718 variants produced a markedly increased MLP when the APOE alleles were included. The largest increase was for rs145449661, from 0.59 to 2.24.

In the analyses not including principal components, the OR was over 2 for 58 variants and under 0.6 for 10. When all 68 of these variants were entered into a ridge regression analysis there were 14 for which \(|\beta/SE(\beta)| \) exceeded 1.3, consisting of rs140824606, rs149626525, rs7258166, rs1477711004, rs116094317, rs41289512, rs142778802, rs79701229, rs6857, rs34404554, rs11556505, rs429358, rs7412 and rs1081105. However in preliminary analyses rs66626994 provided some evidence of being independently associated with LOAD so it was added to the list of variants to be used as covariates in the likelihood ratio tests for each of the 68 individually associated variants. The results of these likelihood ratio tests are in Supplementary Table S3. The main finding of note is that for every variant initially showing strong evidence of association with LOAD the MLP dramatically decreased when rs429358 was included as a covariate. In the adjusted analyses the strongest evidence for independent association was for rs66626994, which produced MLP=3.93 when rs429358 was included and MLP=3.58 when both rs429358 and rs7412 were included. By contrast, the MLP for rs429358 itself remained over 20 no matter which other variants were included as covariates. The MLP for rs7412 was 6.70 on its own and fell to 2.56 when rs429358 was included. The MLPs obtained when both rs429358 and rs7412 were included as covariates were not markedly different from those obtained when including only rs429358.

In terms of effect size, the estimate for the OR for rs429358 on its own was 3.05 with 95% confidence interval 2.65-3.51. Including other variants as covariates had very little impact on this estimate, the lowest value of OR = 2.55 (2.15-3.02) being obtained when including rs6857. Thus there is no suggestion that any other variants in the region have a modifying effect on rs429358. By contrast, there were striking differences in the effect sizes for rs429358 between different ancestry groups, with OR = 5.36 (4.18-6.87) in NHW subjects, OR = 1.72 (1.35-2.20) in CH subjects and OR = 3.15 (2.41-4.11) in AA subjects. Within ancestry groups, these estimates were not impacted by including other variants as covariates.

For rs7412, the unadjusted estimate was OR = 0.53 (0.42-0.68). With rs429358 also included this changed to 0.68 (0.53-0.88) but including other variants had little effect. Including
rs429358 produced OR = 0.62 (0.37-1.05) in NHW subjects, OR = 0.83 (0.55-1.26) in CH subjects and 0.65 (0.42-0.98) for AA subjects. Again, within ancestry groups variants apart from rs429358 had little effect on the estimates.

For rs66626994, the estimate on its own was OR = 2.13 (1.84-2.48). With rs429358 and rs7412 included this fell to OR = 1.37 (1.15-1.62) in the whole cohort and was OR = 1.34 (.95-1.89) in NHW subjects, OR = 1.00 (0.73-1.38) in CH subjects and OR = 1.51 (1.09-2.08) in AA subjects.

Analyses of variants subdivided by rs429358 genotype did not reveal any which showed independent evidence for association with LOAD risk, either in the sample as a whole or in individual ancestry groups.

The results from the various analyses detailed above show that, after correction for multiple testing, none of the common variants tested had a statistically significant independent association with LOAD risk apart from rs429358. Further examination of results broken down by ancestry and by rs429358 genotype did not reveal evidence that other variants were independently exerting an effect within ancestry cohorts.

Discussion
Given the sample size, the results from the weighted burden analyses are consistent with expectations, in that no gene produces statistically significant results once the effect of rs429358 is incorporated. This suggests that the initially significant result obtained for TOMM40 simply results from LD between variants in it and rs429358. The fact that other genes with prior convincing evidence for involvement in LOAD do not produce significant results is unsurprising. When similar methods were applied to a sample approximately five times larger significant results were obtained for TREM2, ABCA7 and SORL1 (Curtis et al., 2019). Of these, the most significant was TREM2, which produced a likelihood ratio statistic of 57.9. Since this is distributed as a chi-squared statistic with one degree of freedom, the expectation for the likelihood ratio statistic to be produced from the current sample is approximately 1+(57.9-1)/5=12.4. In fact, the actual likelihood ratio statistic produced was only slightly lower than this, 12.0, indicating that the current results are entirely consistent with those obtained previously but that the sample size is too small to expect to obtain significant results.

More detailed analysis of variants in the region did not produce any convincing evidence for any other variant to exert a causative effect on LOAD risk. (Here, we use the term causative to mean that a variant itself makes a direct contribution to risk rather than showing association for some other reason.) Of course it is not possible to completely exclude the possibility that one or more of them might exert some small effect which might be detected in a larger sample, either on their own or through epistatic effects on the APOE variants. However for all variants showing initially strong evidence of association with LOAD, the
evidence dramatically diminishes when rs429358 is included as a covariate. This implies that their association is at least primarily driven by LD relationships with rs429358. Conversely, the evidence supporting rs429358 remains very strong no matter which other variants are included as covariates, consistent with the notion that it exerts a direct causative effect. We might note that the current study does not provide support for a protective role for rs7412-T, since the initial MLP was only 6.70 and this fell to 2.56 when rs429358 was included. Using the standard allelic notation, this is equivalent to initially comparing the effects of APOE ε2 against ε3 and ε4 combined and then subsequently comparing ε2 against ε3 alone. The lower risk associated with ε2 against ε3 is a consistent finding of other studies and the relatively weak signal in the current study is expected given the low allele frequency of rs7412-T and should not be taken to undermine support for it having a real protective effect. The very low risk of LOAD in rs7412-T homozygotes has recently been confirmed in a large study (Reiman et al., 2020).

In order to assess whether the current sample would have the power to detect a variant modifying the effect of rs429358 in such a way at to explain the differential effect across ancestries, we should begin by noting that in this heterogenous sample the effect of such a variant should emerge more strongly when the APOE variants are included as covariates. In fact, there is no variant for which the MLP increases substantially. Instead, we tend to observe the opposite effect, explained by variants simply being in LD but having no direct causative effect themselves. To quantify this further, we can consider the situation in which there is a second variant affecting risk which has different LD relationships with rs429358 in different ancestry groups. The effect on differential risk will be maximal if it is in complete LD with rs429358 with reciprocal relationships such that in the NHW subjects the higher risk allele always occurs in phase with the C allele of rs429358 (which confers increased risk) while in CH subjects it is the lower risk allele of the second variant which always occurs in phase with the C allele of rs429358. In this scenario, we observe that the two variants together, tagged by rs429358, have an OR of 5.36 in NHW subjects and 1.72 in CH subjects, equivalent to logistic regression coefficients of the natural logs of these ORs, 1.68 and 0.53. We could then break this down as a contribution from the two loci separately, with rs429358 having a coefficient of 1.105 and the putative second variant having a coefficient of 0.575 to fit the observed values (since 1.105+0.575=1.68 and 1.105-0.575=0.53). This is the minimum effect size which a single variant explaining the differential risk between ancestries could have. If the LD relationships were weaker the effect size of the variant would need to be larger. In an analysis which included rs429358 genotype as a covariate the effect of this variant would emerge and the OR associated with this coefficient of 0.575 would be exp(0.575)=1.8. Using a standard sample size calculator (https://www.stat.ubc.ca/~rollin/stats/ssize/caco.html), we find that in these samples a common variant with OR of 1.8 has over 90% power to produce an MLP of 6 whereas the highest MLP we observe when rs429358 genotype is included as a covariate is only 3.93 (as shown in Supplementary Table 3). From these considerations, we conclude that there is
good power to detect a variant in the region which would explain the differential effect of rs429358 across ancestries and that it is unlikely that such a variant is in fact present.

In this study we do observe that the effect of rs429358 differs between the ancestry groups. However the low magnitude of the effect noted in the CH sample could be a simple consequence of the lower age of controls relative to cases in this sample, since it would be expected that in this situation any signal from any genetic effect on risk would be reduced. There is no obvious explanation for the intermediate effect size observed in the AA subjects and this might represent some real biological phenomenon or might be due to other, unknown mechanisms such as ascertainment procedures or differential cardiovascular mortality. From other studies, the difference in rs429358 effects between ancestries seems to be a fairly consistent finding which, from the current study, cannot be explained by the presence of another causal variant in LD with it. It is difficult to exclude the possibility that a number of variants might cumulatively have some effect, although the present study does not detect any evidence for this. This would be consistent with the previous findings that local ancestry is associated with rs429358 effect size. If the previously reported haplotypes associated with LOAD risk via effects on gene expression of APOE were in different LD relationships with rs429358 in different ancestry groups then this might potentially account for the differential association (Zhou et al., 2019). This could be investigated using large multi-ethnic samples genotyped for the relevant markers.

Theoretically it remains a possibility that there could be an unknown variant which was driving this effect. An example would be the poly-T polymorphism in intron 6 of TOMM40, rs10524523, which is in LD with rs429358 and which has been inconsistently reported to be associated with risk of LOAD and also a large number of other related phenotypes as reviewed recently (Chiba-Falek et al., 2018). This variant was not provided in the data release, presumably because of problems calling a single nucleotide repeat accurately from next generation sequencing pipelines.

When looked at in their entirety, reports of involvement of individual genes and variants in this region resemble the findings which were reported in the candidate gene era. Because of LD relationships with rs429358 they do show a primary association with LOAD. But when studies try to examine whether they exert independent effects one runs into problems with small sample size, subgroup analyses, application to different phenotypes and, presumably, publication bias. We should start from the position that it is inherently unlikely that by chance there will be a gene close to rs429358 which happens to be causative. It could also be argued that it is a priori quite unlikely that variants such as intronic single nucleotide repeats and intergenic variants would have important functional effects. Taking this consideration into account and incorporating the results from the current study it seems reasonable to conclude that, aside from rs429358 and rs7412, other variants in this region do not individually have a substantial effect on risk of LOAD.
References


Acknowledgments

The author wishes to acknowledge the staff supporting the High Performance Computing Cluster, Computer Science Department, University College London. This work was carried out in part using resources provided by BBSRC equipment grant BB/R01356X/1.

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 Disease 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.; Mesoscale 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.

The Alzheimer’s Disease Sequencing Project (ADSP) is comprised of two Alzheimer’s Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer’s Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984),


and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. Seshadri and Fornage. Data generation and harmonization in the Follow-up Phases is supported by U54AG052427 (to Drs. Schellenberg and Wang).

The ADGC cohorts include: Adult Changes in Thought (ACT), the Alzheimer’s Disease Centers (ADC), the Chicago Health and Aging Project (CHAP), the Memory and Aging Project (MAP), Mayo Clinic (MAYO), Mayo Parkinson’s Disease controls, University of Miami, the Multi-Institutional Research in Alzheimer’s Genetic Epidemiology Study (MIRAGE), the National Cell Repository for Alzheimer’s Disease (NCRAD), the National Institute on Aging Late Onset Alzheimer’s Disease Family Study (NIA-LOAD), the Religious Orders Study (ROS), the Texas Alzheimer’s Research and Care Consortium (TARC), Vanderbilt University/Case Western Reserve University (VAN/CWRU), the Washington Heights-Inwood Columbia Aging Project (WHICAP) and the Washington University Sequencing Project (WUSP), the Columbia University Hispanic- Estudio Familiar de Influencia Genetica de Alzheimer (EFIGA), the University of Toronto (UT), and Genetic Differences (GD).

The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193. The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904, the EU Joint Programme - Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435. ARIC research is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899,
2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01-01HL70825 from the NHLBI. CHS research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS (R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their participants, faculty and staff. The content of these manuscripts is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services.

The four LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), The American Genome Center at the Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University Genome Institute (U54HG003079).

Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Cell Repository for Alzheimer’s Disease (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Study Investigators, the NIA funded Alzheimer’s Disease Centers (ADCs), and the National Alzheimer’s Coordinating
Center (NACC, U01AG016976) and the National Institute on Aging Genetics of Alzheimer’s Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA, and at the Database for Genotypes and Phenotypes (dbGaP) funded by NIH. This research was supported in part by the Intramural Research Program of the National Institutes of health, National Library of Medicine. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or nongovernmental organizations.

**Conflict of interest statement**

The author declares no conflict of interest.

**Data availability statement**

Data sharing is not applicable to this article as no new data was created in this study. However scripts, supporting files, interim files and full results will be deposited at the NIAGADS site: https://www.niagads.org/adsp/content/home.
Table 1

Breakdown of samples by ethnicity, age and sex. AA = African American; NHW = Non-Hispanic White; CH = Caribbean Hispanic.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Status</th>
<th>Male:Female</th>
<th>Age mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>Controls</td>
<td>41:160</td>
<td>76.4 (6.5)</td>
</tr>
<tr>
<td></td>
<td>Cases</td>
<td>83:165</td>
<td>72.7 (6.9)</td>
</tr>
<tr>
<td>NHW</td>
<td>Controls</td>
<td>250:394</td>
<td>80.1 (6.4)</td>
</tr>
<tr>
<td></td>
<td>Cases</td>
<td>185:154</td>
<td>73.4 (7.6)</td>
</tr>
<tr>
<td>CH</td>
<td>Controls</td>
<td>133:268</td>
<td>69.3 (6.4)</td>
</tr>
<tr>
<td></td>
<td>Cases</td>
<td>93:177</td>
<td>74.7 (6.8)</td>
</tr>
<tr>
<td>All</td>
<td>Controls</td>
<td>424:822</td>
<td>76.0 (8.0)</td>
</tr>
<tr>
<td></td>
<td>Cases</td>
<td>361:496</td>
<td>73.6 (7.2)</td>
</tr>
</tbody>
</table>
The table shows the MLP for the likelihood ratio test for each variant as a predictor of LOAD including 20 principal components as covariates, with and without also including rs429358 and rs7412. All variants for which the second MLP was greater than 2.5 are shown.

, with different covariates included in the model, for all variants which initially produced and MLP > 20.

<table>
<thead>
<tr>
<th>SNP</th>
<th>rs429358 and rs7412 not included</th>
<th>rs429358 and rs7412 included</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs59007384</td>
<td>18.97</td>
<td>3.72</td>
</tr>
<tr>
<td>rs66626994</td>
<td>21.78</td>
<td>3.66</td>
</tr>
<tr>
<td>rs11556505</td>
<td>23.57</td>
<td>3.2</td>
</tr>
<tr>
<td>rs2075650</td>
<td>21.88</td>
<td>3.11</td>
</tr>
<tr>
<td>rs6857</td>
<td>36.45</td>
<td>2.97</td>
</tr>
<tr>
<td>rs12721046</td>
<td>30.47</td>
<td>2.96</td>
</tr>
<tr>
<td>rs186113697</td>
<td>1.93</td>
<td>2.95</td>
</tr>
<tr>
<td>rs205909</td>
<td>5.71</td>
<td>2.9</td>
</tr>
<tr>
<td>rs34954997</td>
<td>20</td>
<td>2.87</td>
</tr>
<tr>
<td>rs111789331</td>
<td>30.45</td>
<td>2.86</td>
</tr>
<tr>
<td>rs149626525</td>
<td>3.59</td>
<td>2.82</td>
</tr>
<tr>
<td>rs434132</td>
<td>9.17</td>
<td>2.6</td>
</tr>
<tr>
<td>rs157587</td>
<td>3.43</td>
<td>2.6</td>
</tr>
<tr>
<td>rs5117</td>
<td>20.05</td>
<td>2.58</td>
</tr>
</tbody>
</table>