Published in final edited form as:

Ann Neurol. 2015 September ; 78(3): 487–498. doi:10.1002/ana.24466.

Rare coding mutations identified by sequencing of Alzheimer's disease GWAS loci

Badri N. Vardarajan[1,2,3,5,*], **Mahdi Ghani**[8,*], **Amanda Kahn**[1], **Stephanie Sheikh**[1,3], **Christine Sato**[8], **Sandra Barral**[1,2,3], **Joseph H. Lee**[1,2,6], **Rong Cheng**[1,2], **Christiane Reitz**[1,2,3], **Rafael Lantigua**[1,7], **Dolly Reyes-Dumeyer**[1,2], **Martin Medrano**[10], **Ivonne Z. Jimenez-Velazquez**[11], **Ekaterina Rogaeva**[8], **Peter St George-Hyslop**[8,9,**], and **Richard Mayeux**[1,2,3,4,6,**]

[1]The Taub Institute for Research on Alzheimer's Disease and the Aging Brain, College of Physicians and Surgeons, Columbia University, New York, NY, USA

[2]The Gertrude H. Sergievsky Center, College of Physicians and Surgeons, Columbia University, New York, NY, USA

[3]The Department of Neurology, College of Physicians and Surgeons, Columbia University, New York, NY, USA

[4]The Department of Psychiatry, College of Physicians and Surgeons, Columbia University, New York, NY, USA

[5]The Department of Systems Biology, College of Physicians and Surgeons, Columbia University, New York, NY, USA

[6]The Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA

[7]The Department of Medicine, College of Physicians and Surgeons, Columbia University, New York, NY, USA

[8]Tanz Centre for Research in Neurodegenerative Diseases, and Department of Medicine, University of Toronto, Krembil Discovery Tower, 60 Leonard Avenue, Toronto, Ontario, Canada, M5T 2S8

[9]Cambridge Institute for Medical Research, Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK, CB2 0XY

[10]School of Medicine, Pontificia Universidad Catolica Madre y Maestra, Santiago, Dominican Republic

Correspondence: Richard Mayeux, MD, MSc., Taub Institute, 630 West 168th Street, Columbia University, New York, NY 10032, Phone: 212-305-2391, rpm2@columbia.edu Or Peter St George-Hyslop, MD, Cambridge Institute for Medical Research, Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK, CB2 0XY, Phone: 44-1223-763369, phs22@cam.ac.uk.

*These authors contributed equally to the manuscript

**Senior authors contribute equally to the manuscript

Conflict of Interest Disclosure: The authors have nothing to disclose

Author Contributions: Conception and design of the study: R.M, P.S.G.H and E.R.
Data collection: R.M, P.S.G.H, E.R, R.L, D.R.D, M.M and I.J.V
Data Analysis: B.N.V, M.G., C.S, A.K, S.S and R.C
Writing of the final manuscript: B.N.V, R.M, M.G, P.S.G.H, E.R, S.B, J.H.L and C.R

[11]Department of Medicine, Geriatrics Program, School of Medicine, University of Puerto Rico, San Juan, Puerto Rico

Abstract

Objective—To detect rare coding variants underlying loci detected by genome-wide association studies (GWASs) of late-onset Alzheimer's disease (LOAD).

Methods—We conducted targeted sequencing of *ABCA7, BIN1, CD2AP, CLU, CR1, EPHA1, MS4A4A/MS4A6A* and *PICALM* in three independent LOAD cohorts: 176 patients from 124 Caribbean Hispanics families, 120 patients and 33 unaffected individuals from the 129 NIA-LOAD Family Study; and 263 unrelated Canadian individuals of European ancestry (210 sporadic patients and 53 controls). Rare coding variants found in at least two datasets were genotyped in independent groups of ancestry matched controls. Additionally, the Exome Aggregation Consortium (ExAC) was used as a reference dataset for population-based allele frequencies.

Results—Overall we detected a statistically significant 3.1-fold enrichment of the non-synonymous mutations in the Caucasian LOAD cases compared with controls (p=0.002) and no difference in synonymous variants. A stopgain mutation in *ABCA7* (E1769X) and missense mutation in *CD2AP* (T374A) were highly significant in Caucasian LOAD cases, and mutations in *EPHA1* (P460L) and *BIN1* (K358R) were significant in Caribbean Hispanic families with LOAD. The *EPHA1* variant segregated completely in an extended Caribbean Hispanic family and was also nominally significant in the Caucasians. Additionally, *BIN1* (K358R) segregated in two of the six Caribbean Hispanic families where the mutations were discovered.

Interpretation—Targeted sequencing of confirmed GWAS loci revealed an excess burden of deleterious coding mutations in LOAD with the greatest burden observed in *ABCA7* and *BIN1*. Identifying coding variants in LOAD will facilitate the creation of tractable models for investigation of disease related mechanisms and potential therapies.

Keywords

Targeted sequencing; GWAS and rare variants; Alzheimer's disease

Introduction

The first large-scale genome-wide association studies (GWASs) using common single nucleotide polymorphisms (SNPs) identified *CLU, PICALM*, *CR1*, and *BIN1* as late onset Alzheimer's disease (LOAD) susceptibility loci[1–3], which were widely confirmed by others[4,5]. The effect sizes of these genetic associations were much smaller than for *APOE*[2, 6] with odds ratios ranging from 1.16 to 1.20. Follow-up GWASs identified additional LOAD susceptibility variants [7, 8]. While the known function of the genes implicated in these GWAS encode proteins implicating disruptions of lipid metabolism, immune response and endocytosis or intracellular trafficking as potential mechanisms in LOAD, only a handful of disease-associated variants in these genes, such as *SORL1*[9, 10], have been identified.

Surprisingly, targeted exome sequencing of large multiplex pedigrees with LOAD identified mutations in *APP, PSEN1* and *PSEN2*[11, 12], indicating that rare coding sequence variants

even in genes associated with early onset AD may account for a portion of disease risk in LOAD. Also, rare coding sequence variants in *ADAM10*[13], *TREM2*[12, 14] and *PLD3*[15] have been found in patients with LOAD. Because the majority of loci detected by SNP-based GWAS of LOAD have not been investigated for rare coding sequence variants, we conducted targeted sequencing of the top eight genetic loci frequently associated with LOAD[7, 8, 16–18], with the exception of the *CD33* locus which was not well replicated in subsequent large meta-GWAS[18].

Methods

Sample selection

All participants (Table 1) were recruited after providing informed consent and with approval by the relevant institutional review boards. Persons deemed unaffected were required to have had documented cognitive testing and clinical examination to verify their clinical status and diagnosis. Families in which patients had known mutations in *APP*, *PSEN1*, *PSEN2*, *GRN*, or *MAPT* were excluded. All selected probands came from families with four or more affected individuals.

NIA-LOAD/NCRAD Study

Affected and unaffected individuals (n=153) from 129 families within the NIA-LOAD Family Study[5] were selected for targeted sequencing analysis, including 120 individuals with LOAD and 33 similarly aged unaffected (Table 1). Patients had a mean age of onset of 75.1±8.3 years with 38% frequency of the *APOE ε4* allele, and unaffected participants were older (mean age of 82.2±10.8 years) with 13% frequency of the *APOE ε4* allele.

Estudio Familiar de Influencia Genetica en Alzheimer (EFIGA)

Recruitment for the EFIGA family study began in 1998, and was restricted to Caribbean Hispanics[19], mostly from the Dominican Republic. 176 affected patients from 124 families were selected for targeted sequencing (the mean age of onset was 74.8±8.3 years; and 63.1% were woman).

Toronto LOAD Study

Targeted sequencing of GWAS loci was conducted in 210 well-characterized sporadic LOAD patients and 53 normal controls of European ancestry from the GenADA study based on sufficient quantity/quality DNA samples. The mean age of onset in cases was 73.9±7.3 years, 50.4% were women and *APOE ε4* allele frequency was 35.7%. These patients were either clinically diagnosed with probable LOAD (n=169) or autopsy-confirmed LOAD cases from the brain bank at the Tanz Center for Neurodegenerative Research in Toronto (n=41). Mean age at the time of examination in controls was 80.3±3.6 years, 62.3% of them were women, and the *APOE ε4* allele frequency was 22.4%.

Exome Aggregation Consortium (ExAC: http://exac.broadinstitute.org)

The ExAC dataset was used as a reference dataset of population-based allele frequencies. It contains data from 60,706 unrelated adult individuals sequenced as part of various disease-

specific and population genetic studies from six different ethnic groups. LOAD was not one of the diseases investigated. We used the Non-Finnish European and Latino/American cohorts to compare the frequencies of variants found in Caucasian and Caribbean Hispanic cohorts respectively.

Sample Preparation

High molecular weight DNA was isolated from either fresh or frozen samples that had been stored at −80°C. Blood genomic DNA was isolated using the Gentra Puregene and FlexiGene kits (Qiagen), and saliva genomic DNA was isolated using the prepIT.L2P (DNAgenoteck Inc). When high quality DNA derived from blood was unavailable, lymphocyte cell lines were used (in a total of 13 probands). DNA concentration was determined by nanodrop for most analyses.

Targeted Sequencing

Target enrichment of the samples was performed using the Agilent SureSelect system (for Hispanics and NIA LOAD dataset), and Roche NimbleGen SeqCap EZ Designs-custom (for Toronto dataset). Custom oligonucleotide baits captured exonic regions and splice sites of the genes of interest and amplified. For the SeqCap EZ approach, the sequencing library was hybridized to the SeqCap EZ Oligo pool that was made against the target regions of interest. The end product was subjected to high throughput sequencing. After the DNA samples were prepared, they were multiplexed with index “barcode” primers and pooled for sequencing in batches of up to 12 samples.

Sequencing of all samples occurred on Illumina’s Genome Analyzer IIx, HiSeq 2000, and MiSeq platforms (http://www.illumina.com). Paired-end reads were performed over 82–307 sequencing cycles. Data files were demultiplexed by “barcode” to separate pooled samples into individual probands. We were able to obtain high coverage at an average depth of >1000× per sample and interval region captured.

Follow-up Genotyping

Caribbean Hispanics with mutations also observed in one of the other two datasets underwent genotyping or Sanger sequencing to confirm non-synonymous variants. To determine the population frequencies for variants discovered within our datasets, we genotyped unrelated controls of the same ethnic background (Table 1). We also conducted validation genotyping in 13 Caribbean Hispanic families (n=148) of the patients where the variants were discovered. Additionally, to compare the allele frequencies for novel variants identified in this study from unaffected persons in the Caribbean Hispanic population, we genotyped 460 unaffected and unrelated persons (68.0% women, mean age at examination was 81.2±7.1, and *APOE ε4* allele frequency was 12.9 %) of the same ancestry. We also genotyped 444 white, non-Hispanic controls (58.0% women, the mean age at the time of examination was 84.7±5.5 years, and *APOE ε4* allele frequency was 10.0%) in the NIA-LOAD dataset in order to estimate population frequencies for the mutations discovered. The controls were determined to be of the same ethnic background as the familial cases using methods described previously[19]. Follow-up genotyping was also done on 238 normal controls matched to the Toronto sporadic LOAD dataset by ethnic origin, sex and age

(57.5% were women, the mean age at the time of examination was 73.1±9.4 years, and *APOE ε4* allele frequency was 14.4%). Genotypes were generated using SEQUENOM's MassArray iPLEX technology, following the manufacturers' instructions. The system involves multiplex PCR and mini-sequencing assays, followed by MALDI-TOF mass spectrometry analysis.

Analytical Methods

We aligned the reads obtained from the pooled sequencing to the human reference genome build 37 using the Burrows Wheeler Aligner[20] (http://bio-bwa.sourceforge.net/). Quality control of the sequencing data was done using established methods, including base alignment quality calibration and refinement of local alignment around putative indels using the Genome Analysis Toolkit (GATK)[21]. Variants were called and recalibrated using multi-sample calling with GATK's UnifiedGenotyper and VariantRecalibrator modules. Reliably called variants were annotated by ANNOVAR[22] including *in-silico* functional prediction using POLYPHEN[23] and extent of cross-species conservation using PHYLOP[24].

Burden Tests: We estimated the burden of different classes of mutations (loss of function, all non-synonymous and all synonymous variants) using a binomial test as described here[25]. To determine if a class of mutations was enriched in cases, we used a binomial test with probability of success equal to the frequency of mutation class in controls (background or expected frequency). Also, any bias introduced in the test due to an unbalanced case-control set was compared to observations from the synonymous mutation class that was used to set the background expectation.

Individual SNVs significance tests: To test the association of individual SNVs with LOAD, we compared the allele frequencies of SNVs in patients with unaffected samples from follow-up genotyping combined with the publicly available Exome Aggregation Consortium (ExAC: http://exac.broadinstitute.org) data using Fisher's exact test. We used this dataset to provide a much larger and more representative estimate of allele frequencies than could be ascertained from the smaller NIA-LOAD and Toronto datasets alone. Because of the lack of an optimal ethnically matched control dataset for Caribbean Hispanics, we used the Latino American cohort for an estimate of allele frequencies of rare variants. Additionally for the Caribbean Hispanic cohort only, we tested segregation and LOAD association in this dataset using Generalized Estimation Equations (GEE) to adjust for familial correlation.

Results

Sequencing

We identified 12 coding mutations in seven genes in at least two of the three datasets, including seven autopsy confirmed LOAD cases (Table 2). These twelve coding mutations included: four mutations in *ABCA7*, two each in *CD2AP* and *PICALM* and one each in *BIN1*, *CLU*, *EPHA1* and *MS4A6A*. Three rare coding mutations were observed in cases from all three datasets: rs138047593 in *BIN1*, rs202178565 in *EPHA1*, and rs138650483 in *MS4A6A*, the *EPHA1* and *BIN1* mutations were subsequently confirmed by follow-up genotyping in Hispanic cohort. We assessed the association of these variants independently

in Caucasian and Hispanic cohorts by comparing them against the population-based allele frequencies available in the ExAC database and by testing family-based association in Caribbean Hispanic families.

Caucasians

For the 12 variants detected in NIA-LOAD and Toronto datasets (Table 3), we compared the frequency of SNVs between 330 cases in these datasets with the 33,370 non-Finnish Europeans from ExAC using a Fisher's exact test. A stopgain mutation in *ABCA7* (E1769X) and missense mutation in *CD2AP* (T374A) were highly significant after correction for multiple testing (p=5.3e-04 and 5.3e-08 respectively). Of the remaining variants discovered in multiple datasets, one rare variant in both *EPHA1* and *PICALM* were nominally significant (p=0.03 and 0.007 respectively). The p.K358R variant in *BIN1* was observed in 1.8% of the ExAC database Caucasians which is similar to the frequency we observed in the cohort of patients here with LOAD. The remaining variants were extremely rare (MAF<0.5%) in the ExAC database Caucasians.

Caribbean Hispanics

For the seven variants found in this dataset (and at least one other Caucasian dataset) (Table 4), we tested segregation and LOAD association using validation-genotyping data in 13 families and 460 independent case-controls. Further, we compared the frequency of the variants in LOAD patients with the Latino allele frequencies (n=5789) in the ExAC database. The p.P460L in *EPHA1* and p.K358R in *BIN1* were significantly associated with LOAD when compared to both internally genotyped Caribbean Hispanic controls and population Latino controls in the ExaC database after correction for multiple testing (Table 4). Notably, the *EPHA1* rs202178565 variant (P460L) was observed in only one of the 490 unaffected Caribbean Hispanic individuals and none of the Caucasian controls (Table 4). This *EPHA1* mutation also segregated completely in four affected members of a Caribbean Hispanic family (Figure 1). The variant was significant both in the Fisher exact test (p= 2.6e-03) and regression model (p=8.64e-05) in Caribbean Hispanics and nominally significant in the Caucasian cohort (p=0.03).

Follow-up genotyping of the *BIN1* p.K358R mutation revealed that it was predominantly found in affected members with LOAD from six Caribbean Hispanics families. We observed *BIN1* p.K358R in 11 LOAD patients and only three elderly controls (over 65 years) in these families. We also observed the mutation in nine unaffected members under the age of 65 years (average age=54 years). We observed a higher frequency of the mutation in the families (0.085 in familial cases and 0.069 in familial controls) compared to genotyped Caribbean Hispanic controls (0.0084) and Latino population controls from the ExAC database (0.0026). This variant was significantly associated with Caribbean Hispanic LOAD families in both a regression model (p=1.27e-05) and Fisher's exact test (p=5.85e-04 The *BIN1* p.K358R allele frequencies in Caucasian and Caribbean Hispanic population controls were similar. However, we found much higher frequency of this variant in families suggesting that the effect of this variant in multiplex families may be due to epistasis with other genetic or environmental risk factors. Further investigation of this mutation is required to evaluate the effect of this variant in LOAD pathogenesis.

Other mutations

In addition to mutations observed in multiple datasets, a total of 88 rare damaging mutations were found to present in individual datasets and only detected in patients with LOAD: 21 in NIA-LOAD, 37 in Toronto and 30 in the Caribbean Hispanics. When compared to the ExAC population frequencies, 38 out of 88 variants were nominally significant at $p<0.05$ (Supplementary Table 1), 21 of which were observed in *ABCA7* and five in *EPHA1*. All the nominally significant variants were extremely rare in the general population (max MAF= 0.05%) and a majority of them were predicted to be deleterious to the coding protein.

Burden Tests

We calculated the overall burden of these novel or rare coding non-synonymous mutations (including SNVs and short indels) compared with the burden of synonymous mutations in cases and controls for each gene in the three datasets (Table 5). Combining the observations from the NIA-LOAD and Toronto Caucasian datasets, we detected a statistically significant 3.1-fold enrichment of the non-synonymous mutations in cases versus controls ($p=0.002$). The LOAD cases also carried 2.76 times more loss of function mutations (stop-loss, stop-gain, frameshift or splicing) and damaging missense mutations, compared to controls ($p=0.02$). In contrast, we did not observe a difference in synonymous mutations for LOAD cases in the two Caucasian datasets compared with controls (1.07 fold, $p=0.59$). The mutation rate per Caribbean Hispanic LOAD patients was comparable to that in the Caucasian dataset across all genes (Table 5).

In total, 11.1% of all patients with LOAD from three datasets were carriers of at least one coding *ABCA7* mutation. Remarkably, 47% of all potentially damaging mutations were observed in the *ABCA7* gene. Of the rare mutations, 8% were detected in *EPHA1* affecting 3.1% of investigated LOAD cases and only a single Caribbean Hispanic control. These results are striking because based on a recent study[26] of thousands of exomes, *ABCA7* and *EPHA1* are highly conserved genes and ranked in the top second and eleventh percentiles, respectively, for intolerance towards mutation in the general population. The high mutation rate in LOAD compared to controls in the highly conserved *ABCA7* and *EPHA1* implies a putative functional role in the pathogenesis of LOAD.

BIN1 was strong contributor to the increased mutation rate in cases compared to controls showing damaging variants in 19 cases (3.75%) but none in controls (Table 2 and supplemental Table 1). The most frequent mutation in the patients was in *BIN1* (p.K358R) where we identified carriers in eight Caucasian (including four autopsy cases) and six Hispanic patients.

There is prior evidence of increased expression of *ABCA7*, *BIN1*, *MS4A6A* in LOAD brains [27] and increased *ABCA7* expression is associated with clinical dementia rating (CDR)[28], with higher expression being associated with more advanced cognitive decline. *BIN1* expression levels were associated with disease progression, where higher expression was associated with a delayed age at onset. However there was no evidence of differential expression of EPHA1 in LOAD compared with controls[28].

Discussion

The results presented here imply that the loci from GWAS associated with LOAD likely contain multiple rare, damaging mutations that can be recurrent among unrelated patients and in some instances, can segregate within families. The dense coverage we used for targeted sequencing allowed for the identification of variants that might not have been detectable with more sparse coverage used in current whole exome or whole genome approaches. Despite the observation that variants in *BIN1*(p.K358R), *EPHA1* (p.P460L) and *MS4A6A* (p.V218M) were found in patients with LOAD from all three datasets, we could not establish statistical significance of the findings due to the rarity of the mutations. However, in the two Caucasian datasets we found statistically significant variants in *CD2AP* and ABCA7, while in the Caribbean Hispanic dataset statistically significant variants were found in *EPHA1* and *BIN1*. The nominally significant variants from individual datasets (Supplementary Table 1) were observed in the ExAC dataset at very low frequencies, providing further support that greater depth of targeted sequencing allows identification of very rare events.

In three datasets enriched by families multiply affected with LOAD, we sequenced eight GWAS loci with consistent SNP-based associations with LOAD across multiple investigations[18]. Analysis of two Caucasian datasets revealed a significantly greater burden of rare and novel non-synonymous (including SNVs and indels) alterations (p=0.002) in cases compared to controls, while the mutation rate of synonymous variants was the similar in cases and contorls. In LOAD we also observed a significant (p=0.02) three-fold enrichment in the subset of alterations that were predicted to be damaging (by POLYPHEN or SIFT).

The greatest burden of damaging sequence variants was found in *ABCA7.* Among Caucasians LOAD cases, we detected 39 carriers of rare variants (20 in NIA-LOAD and 18 in Toronto dataset), constituting 11.8% of 330 investigated cases, while only one carrier of such a variant was found among the 86 sequenced controls (1.2%) (Table 2 and supplemental Table 1). In addition to non-synonymous *ABCA7* variants, we observed a splice site, a stop mutation and frameshift deletions, suggesting a loss-of-function mechanism associated with LOAD. Indeed, our recent functional studies of ABCA7 strongly support such a possibility[29, 30], since suppression of ABCA7 *in vitro* and *in vivo* resulted in an elevation of amyloid production. The complex function of ABCA7 includes mediation of the biogenesis of high-density lipoprotein with cellular lipid and helical apolipoproteins[31], as well as function in apolipoprotein-mediated phospholipid and cholesterol efflux from cells.[32] Finally, a direct role of ABCA7 in APP processing may be associated with its primary biological function to regulate endocytic pathways[30]. Importantly, we previously identified *ABCA7* as a major genetic risk LOAD locus in the African Americans[33], and a whole-genome sequencing study in a large Icelandic cohort identified excess burden of rare loss of function variants in *ABCA7* in LOAD[34]. We confirmed two *ABCA7* loss of function variants reported in that study (c.4416+2T>G and p.Leu1403Argfs*7) and discovered three additional variants (p.708_710del, p.R1489X and E1679X). Our analyses confirm that *ABCA7* has the highest burden or deleterious variants in LOAD, but differences in the

observed mutations could be due to ethnicity, capture and coverage differences in the two studies.

BIN1 was also strongly associated in the burden analysis, with damaging variants in 17 cases (5.1%) but no controls. Several SNPs upstream of the *BIN1* locus have been identified in different GWASs with the largest effect sizes after *APOE* (e.g. rs6733839 with population attributable fraction of 8.1% [35]). *BIN1* transcript levels were increased among LOAD brains compared to controls [36], but coding mutations have not been widely explored. So far, there are only four *BIN1* coding variants with clinical significance listed in the ClinVar database (p.K575*, p.R154Q, p. D151N and p.K35N) and all were reported under Autosomal recessive centronuclear myopathy. Recently, Tan et al. reported that a novel *BIN1* missense mutation p.P318L among the Han Chinese could increase risk of developing AD[37], which was not detected in our datasets. The *BIN1* mutations reported here included p.K358R, identified in eight Caucasian and six Hispanic LOAD patients, as well as p.S267L and p.S202T, each identified in a single LOAD patient. None of these mutations were found in controls or unaffected family members. We observed a strong association between LOAD and *BIN1* p.K358R only in the Caribbean Hispanics. The allele frequency of this variant in the Caucasian patients was similar to the general population. *BIN1* p.K358R is a good candidate for functional studies based on its relatively high frequency in familial LOAD cases and segregation in Caribbean Hispanic families. Importantly, *BIN1* p.K358R likely contributes to LOAD independently from the GWAS SNPs, since it is mapped to a different LD block (Figure 2).

We also identified six non-synonymous variants in *EPHA1*, including p.H888Y, p.R791H, p.V514I, p.R471Q, p.P460L and p.R337Q. The damaging *EPHA1* variant p.P460L (rs202178565) was identified in cases in all three datasets and was absent among our controls as well as in 1000 Genomes and ExAC server dataset. This variant segregated with the LOAD in a Caribbean Hispanic family from the Dominican Republic (Figure 1), supporting its causative role. The *EPHA1* p.P460 amino acid is highly conserved in all mammals and predicted to have a damaging effect on the protein by POLYPHEN estimation. However, the biological impact of this mutation remains to be investigated because there is only limited information on the function of protein. Ephrin receptor A1 encoded by *EPHA1* belongs to the ephrin receptor subfamily of the protein-tyrosine kinase family and plays roles in cell and axonal guidance and synaptic plasticity.

A rare variant was found in *MS4A6A*, which affects splicing of one transcript of the gene (NM_152852: exon8: c.651+1G>A) and is a missense mutation in another transcript (NM_022349: exon6: c.G652A: p.V218M). The *MS4A6A* p.V218M variant was detected in a single unaffected Caucasian. *MS4A6A* is located among several genes at Chr11q12 that all are associated with the inflammatory response. *MS4A6E* mRNA expression and a SNP nearby the gene (rs670139) are associated with more advanced Braak stages of tangle and plaques in AD brain tissue[28]. However, until now a functional variant in this region has not been identified and the current study might provide the first clue[38].

We identified other rare damaging variants among LOAD associated genes, including *CD2AP* (p.I104N, p.R403G, p.L487V, p.M496I, p.S623N and p.K633R) and *CLU*

(p.V434M). *CD2AP* is an adaptor molecule involved in dynamic actin remodeling and membrane trafficking and *CLU* encodes Clusterin, which is a molecular chaperone[39] and is present in senile plaques, and has been shown to modulate Aβ oligomer assembly [40]. We previously reported rare SNPs and small structural variants within the *CLU* gene that were associated with LOAD[41].

Taken together, the results here imply that multiple rare coding mutations are present in genes identified as LOAD associated GWAS loci. Common variants identified in GWAS frequently occur in non-coding sequences within or between genes, and as a result, their functional relationship to disease risk is often hard to define. The data reported here reveal that GWAS loci could harbor both rare damaging variants and common noncoding variants that are independently associated with LOAD (e.g. in *CLU*)[41]. Thus, targeted sequencing within GWAS loci may enable the discovery of coding variants underlying or contributing to the association with LOAD. The use of non-coding variants to build cellular and animal models of disease is confounded by uncertainties surrounding the temporal- and cell type-specific effects of these non-coding variants on the regulation of gene expression. By contrast, disease-associated coding sequence variants can be used to build facile, tractable cellular and animal models by a variety of simple methods including both standard transgenesis and CRISPR-CAS based methods. Such models can be used to investigate the underlying molecular mechanisms of these genes in the pathogenesis of LOAD.

The individual effect of these rare variants is expected to be small and different variants are likely to be causal in different patients and families. For example, the p.K538R variant in *BIN1*, has a strong effect in the Hispanic families but was not associated with LOAD in the Caucasian cohort. It is likely that such variants confer modified risk of disease or depend on other interacting genes or environmental factors. Identification of such rare coding variants could thus aid in understanding the biology of the disease.

The strengths of this study are the three independent cohorts and the careful phenotyping. The fact that some of the same mutations were observed in two or three of the cohorts adds validity to our observations. While there appears to be increased expression associated with some of the genes containing mutations, further studies are required to examine mutation specific expression and to understand the mechanisms by which these mutations lead to disease.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants from the National Institute of Health and the National Institute on Aging: R37AG15473, P01AG07232, R01AG041797, U24AG026395 (RM), Canadian Institutes of Health Research (ER, PSH), Wellcome Trust, Medical Research Council, Ontario Research Fund and Alzheimer Society of Ontario (PSH) and U24AG21886 for the NIA LOAD Cohort.

References

1. Harold D, Abraham R, Hollingworth P, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet. 2009 Oct; 41(10):1088–93. [PubMed: 19734902]
2. Lambert JC, Heath S, Even G, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet. 2009 Oct; 41(10):1094–9. [PubMed: 19734903]
3. Seshadri S, Fitzpatrick AL, Ikram MA, et al. Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA. 2010 May 12; 303(18):1832–40. [PubMed: 20460622]
4. Jun G, Naj AC, Beecham GW, et al. Meta-analysis Confirms CR1, CLU, and PICALM as Alzheimer Disease Risk Loci and Reveals Interactions With APOE Genotypes. Archives of neurology. 2010 Aug 9.
5. Wijsman EM, Pankratz ND, Choi Y, et al. Genome-wide association of familial late-onset Alzheimer's disease replicates BIN1 and CLU and nominates CUGBP2 in interaction with APOE. PLoS genetics. 2011 Feb.7(2):e1001308. [PubMed: 21379329]
6. Farrer LA, Cupples LA, Haines JL, 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. 1997 Oct 22–29; 278(16):1349–56. [PubMed: 9343467]
7. Naj AC, Jun G, Beecham GW, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet. 2011 May; 43(5):436–41. [PubMed: 21460841]
8. Hollingworth P, Harold D, Sims R, et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet. 2011 May; 43(5): 429–35. [PubMed: 21460840]
9. Pottier C, Hannequin D, Coutant S, et al. High frequency of potentially pathogenic SORL1 mutations in autosomal dominant early-onset Alzheimer disease. Molecular psychiatry. 2012 Sep; 17(9):875–9. [PubMed: 22472873]
10. Vardarajan BN, Zhang Y, Lee JH, et al. Coding mutations in SORL1 and Alzheimer's disease. Annals of neurology. 2014 Nov 7.
11. Lee JH, Kahn A, Cheng R, et al. Disease-related mutations among Caribbean Hispanics with familial dementia. Molecular genetics & genomic medicine. 2014 Sep; 2(5):430–7. [PubMed: 25333068]
12. Guerreiro R, Wojtas A, Bras J, et al. TREM2 Variants in Alzheimer's Disease. The New England journal of medicine. 2012 Nov 14.
13. Kim M, Suh J, Romano D, et al. Potential late-onset Alzheimer's disease-associated mutations in the ADAM10 gene attenuate {alpha}-secretase activity. Hum Mol Genet. 2009 Oct 15; 18(20): 3987–96. [PubMed: 19608551]
14. Jonsson T, Stefansson H, Ph DS, et al. Variant of TREM2 Associated with the Risk of Alzheimer's Disease. The New England journal of medicine. 2012 Nov 14.
15. Cruchaga C, Karch CM, Jin SC, et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature. 2014 Jan 23; 505(7484):550–4. [PubMed: 24336208]
16. Seshadri S, Fitzpatrick AL, Ikram MA, et al. Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA. 2010 May 12; 303(18):1832–40. [PubMed: 20460622]
17. Harold D, Abraham R, Hollingworth P, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet. 2009 Oct; 41(10):1088–93. [PubMed: 19734902]
18. Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013 Dec; 45(12):1452–8. [PubMed: 24162737]
19. Lee JH, Cheng R, Barral S, et al. Identification of novel loci for Alzheimer disease and replication of CLU, PICALM, and BIN1 in Caribbean Hispanic individuals. Archives of neurology. 2011 Mar; 68(3):320–8. [PubMed: 21059989]

20. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15; 25(14):1754–60. [PubMed: 19451168]
21. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010 Sep; 20(9):1297–303. [PubMed: 20644199]
22. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010 Sep.38(16):e164. [PubMed: 20601685]
23. Adzhubei, I.; Jordan, DM.; Sunyaev, SR. Predicting functional effect of human missense mutations using PolyPhen-2. In: Haines, Jonathan L., et al., editors. Current protocols in human genetics. Vol. Chapter 7. 2013 Jan. p. 20
24. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome research. 2010 Jan; 20(1):110–21. [PubMed: 19858363]
25. Neale BM, Rivas MA, Voight BF, et al. Testing for an unusual distribution of rare variants. PLoS genetics. 2011 Mar.7(3):e1001322. [PubMed: 21408211]
26. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS genetics. 2013; 9(8):e1003709. [PubMed: 23990802]
27. Myers AJ, Gibbs JR, Webster JA, et al. A survey of genetic human cortical gene expression. Nat Genet. 2007 Dec; 39(12):1494–9. [PubMed: 17982457]
28. Karch CM, Jeng AT, Nowotny P, Cady J, Cruchaga C, Goate AM. Expression of novel Alzheimer's disease risk genes in control and Alzheimer's disease brains. PloS one. 2012; 7(11):e50976. [PubMed: 23226438]
29. Bamji-Mirza M, Najem D, Walker D, et al. Genetic variations in ABCA7 can increase secreted levels of Aβ40/42 and ABCA7 transcription in cell culture models. Neurobiology of aging. In press.
30. Kanayo S, Sumiko A-D, Shinji Y, Hazrati LN, George-Hyslop PS, Fraser P. ABCA7 Loss of Function Alters Alzheimer Amyloid Processing.
31. Tanaka N, Abe-Dohmae S, Iwamoto N, Yokoyama S. Roles of ATP-binding cassette transporter A7 in cholesterol homeostasis and host defense system. J Atheroscler Thromb. 2011; 18(4):274–81. [PubMed: 21173549]
32. Chan SL, Kim WS, Kwok JB, et al. ATP-binding cassette transporter A7 regulates processing of amyloid precursor protein in vitro. J Neurochem. 2008 Jul; 106(2):793–804. [PubMed: 18429932]
33. Reitz C, Jun G, Naj A, et al. Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E 4, and the risk of late-onset Alzheimer disease in African Americans. JAMA. 2013 Apr 10; 309(14):1483–92. [PubMed: 23571587]
34. Steinberg S, Stefansson H, Jonsson T, et al. Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. Nat Genet. 2015 Mar 25.
35. Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013 Dec; 45(12):1452–8. [PubMed: 24162737]
36. Chapuis J, Hansmannel F, Gistelinck M, et al. Increased expression of BIN1 mediates Alzheimer genetic risk by modulating tau pathology. Mol Psychiatry. 2013 Nov; 18(11):1225–34. [PubMed: 23399914]
37. Tan MS, Yu JT, Jiang T, Zhu XC, Guan HS, Tan L. Genetic variation in BIN1 gene and Alzheimer's disease risk in Han Chinese individuals. Neurobiol Aging. 2014 Jul; 35(7):1781.e1–8. [PubMed: 24582639]
38. Omenn GS, Guan Y, Menon R. A new class of protein cancer biomarker candidates: differentially expressed splice variants of ERBB2 (HER2/neu) and ERBB1 (EGFR) in breast cancer cell lines. J Proteomics. 2014 Jul 31.107:103–12. [PubMed: 24802673]
39. Wyatt A, Yerbury J, Poon S, Dabbs R, Wilson M. Chapter 6: The chaperone action of Clusterin and its putative role in quality control of extracellular protein folding. Advances in cancer research. 2009; 104:89–114. [PubMed: 19878774]

40. Narayan P, Orte A, Clarke RW, et al. The extracellular chaperone clusterin sequesters oligomeric forms of the amyloid-beta(1–40) peptide. Nature structural & molecular biology. 2012 Jan; 19(1): 79–83.

41. Bettens K, Brouwers N, Engelborghs S, et al. Both common variations and rare non-synonymous substitutions and small insertion/deletions in CLU are associated with increased Alzheimer risk. Molecular neurodegeneration. 2012; 7:3. [PubMed: 22248099]

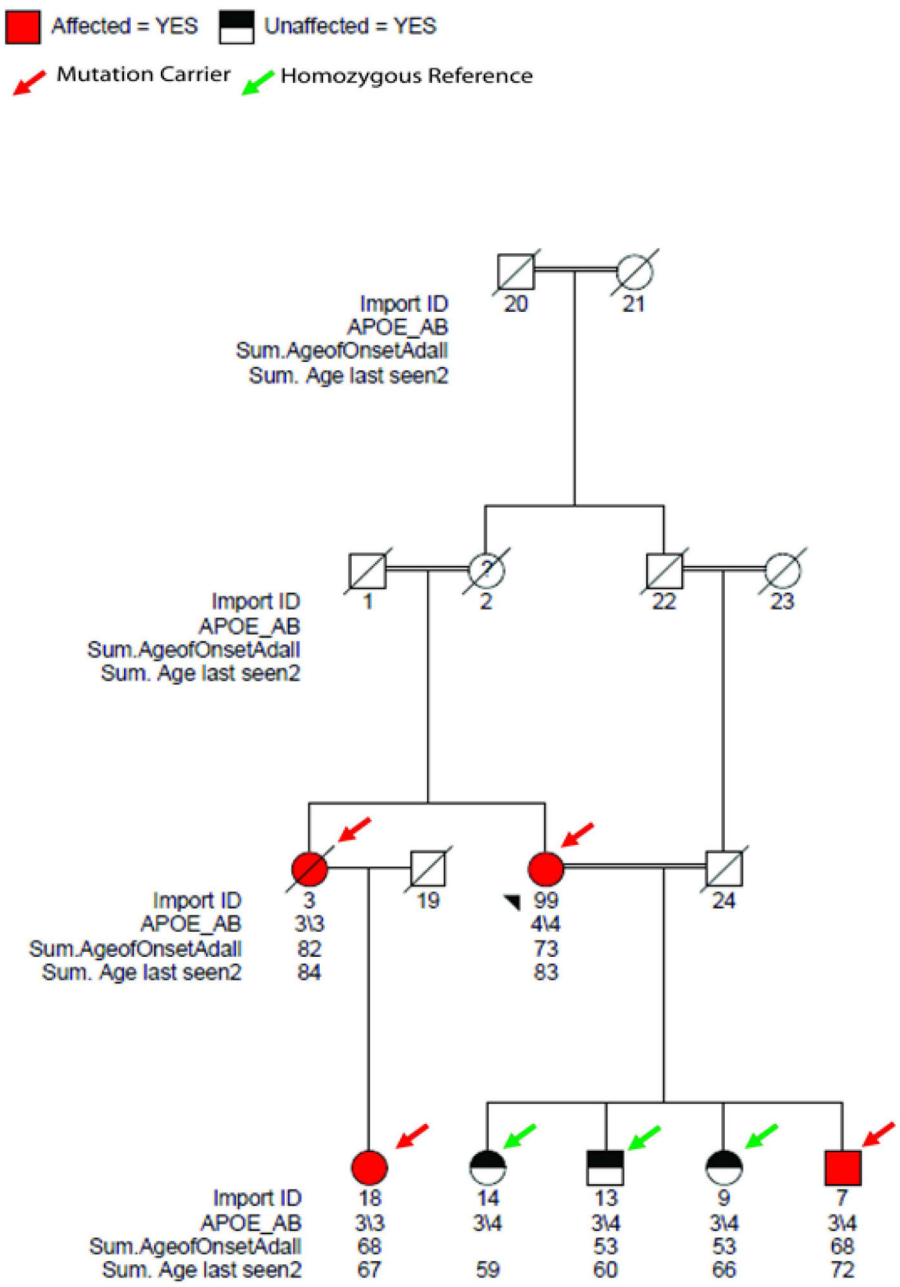

Figure 1.

Missense damaging mutation rs202178565 in EPHA1 (ephrin type-A receptor 1). This mutation was not found in any external controls

Import ID: Internal Subject ID, APOE_AB: APOE ε4 status, Sum.AgeofOnsetAdAll: Age at onset of disease, Sum.Age last seen2: Age of the last examination of the subject

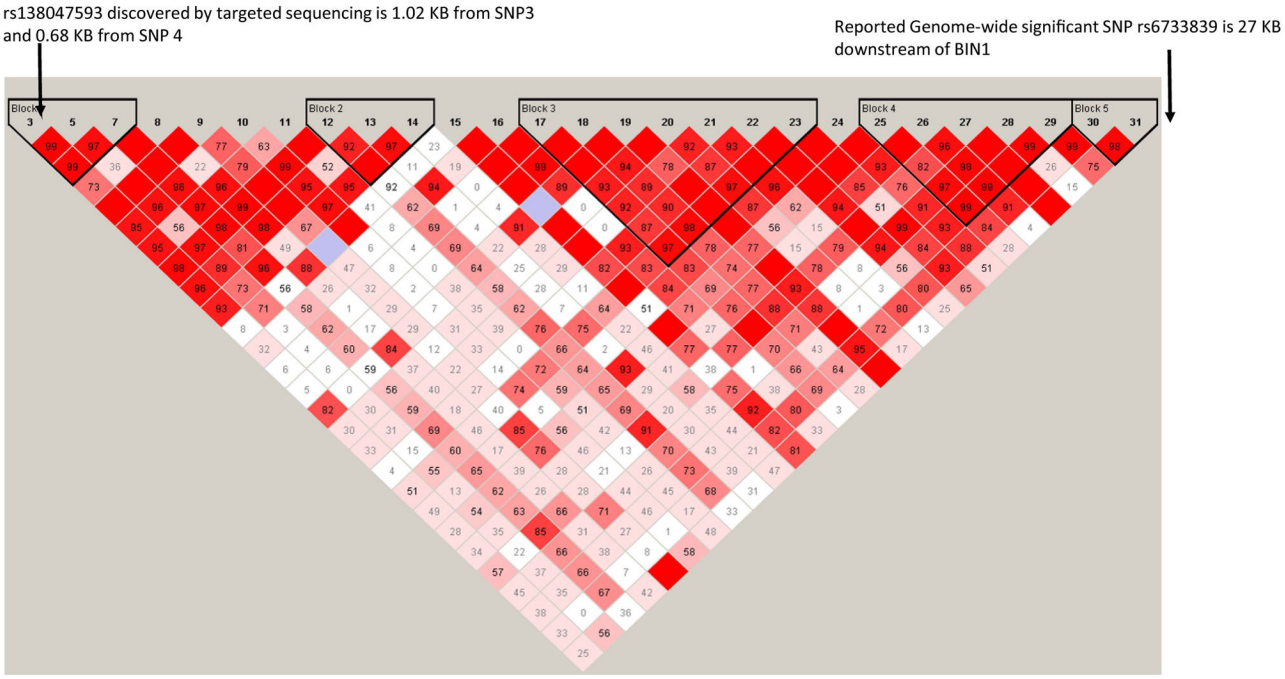

Figure 2.

LD plot of *BIN1* in Hispanics. The LD Plot is generated using 32 genotyped SNPs in 1675 elderly subjects of Caribbean Hispanic ancestry. The reported genome-wide significant hit in Lambert et al (rs6733839) is 27.1 KB upstream of *BIN1*.

Table 1

Demographics of the samples in the targeted sequencing experiment

		Status	Number	Mean Age at Onset or last examination: years ± SD	Women: n (%)	*APOE e4*, %
TARGETED SEQUENCING	**NIA-LOAD**	Affecteds	120	75.1 ± 8.3	77 (64.2)	13.6
		Unaffecteds	33	82.2 ± 10.8	24 (72.7)	34.1
	Toronto	Affecteds	210	73.9 ± 7.3	106 (50.4)	35.7
		Unaffecteds	53	80.3 ± 3.6	33 (62.3)	22.4
	HISPANICS	Affecteds	176	74.8 ± 8.3	111 (63.1)	25.3
FOLLOW-UP GENOTYPING	**HISPANICS**	Unaffecteds	300	81.2 ± 7.1	302 (68.0)	12.9
	WHICAP	Unaffecteds	444	84.7 ± 5.5	174 (58.0)	10.0
	Toronto	Unaffecteds	238	73.1 ± 9.4	137 (57.5)	14.4

Table 2

Annotation of rare or novel non-synonymous SNPS found in at least two of the three datasets

CHR	Position	ID	Ref	Alt	Function	Gene	AA change	POLYPHEN	SIFT
2	127808046	rs138047593*	T	C	nonsynonymous	*BIN1*	K358R	D	D
6	47563608	rs138727736	A	G	nonsynonymous	*CD2AP*	T374A	B	B
6	47591941	rs116754410	A	G	nonsynonymous	*CD2AP*	K633R	D	D
7	143095499	rs202178565**	G	A	nonsynonymous	*EPHA1*	P460L	D	P
8	27462662	rs41276297	G	A	nonsynonymous	*CLU*	T203I	B	B
11	59940500	rs138650483*	C	T	exonic/splicing	*MS4A6A*	V218M	D	D
11	85687719	rs147556602	G	C	nonsynonymous	*PICALM*	P495A	D	D
11	85701307	rs117411388	T	C	nonsynonymous	*PICALM*	H458R	D	D
19	1041971	rs201665195	T	G	nonsynonymous	*ABCA7*	L101R	D	D
19	1051006	rs143718918	G	A	nonsynonymous	*ABCA7*	R880Q	D	D
19	1057343	rs117187003	G	A	nonsynonymous	*ABCA7*	V1599M	D	D
19	1058154	novel	G	T	stopgain	*ABCA7*	E1679X	.	.

* found in all three datasets

** found in all three datasets, not found any unaffected in follow-up genotyping

NS: non-synonymous SNVs

Table 3

Allele Frequency and Fisher tests of SNPs in Caucasians

Gene	ID	TARGETED SEQUENCING (carriers)				TARGETED SEQUENCING FREQ		ExAC FREQ IN EUROPEAN	FISHER TEST P VALUE
		NIA-LOAD		Toronto (autopsy cases)		NIA-LOAD	Toronto		
		UNAFF	LOAD	UNAFF	LOAD	LOAD	LOAD		
BIN1	rs138047593	0	1	0	7 (4)	0.004386	0.016667	1.82E-02	3.69E-01
CD2AP	rs138727736	0	0	0	4 (2)	0	0.009524	4.73E-03	1.37E-01
CD2AP	rs116754410	0	0	0	1 (1)	0	0.002381	3.06E-05	5.33E-08
EPHA1	rs202178565	0	1	0	1	0.004386	0.002381	4.05E-04	3.07E-02
CLU	rs41276297	0	0	0	2	0	0.004762	2.51E-03	2.82E-01
MS4A6A	rs138650483	1	1	0	1	0.00431	0.002381	3.76E-03	1.00E+00
PICALM	rs147556602	0	0	1	0	0	0	3.61E-04	1.00E+00
PICALM	rs117411388	0	2	0	2	0.00885	0.004762	1.11E-03	6.84E-03
ABCA7	rs201665195	0	1	0	1	0.004348	0.002381	1.14E-03	1.82E-01
ABCA7	rs143718918	0	1	0	1	0.004425	0.002381	2.11E-03	3.89E-01
ABCA7	rs117187003	0	4	1	1	0.017857	0.002381	4.16E-03	2.01E-01
ABCA7	19:1058154	0	1	0	1	0.004425	0.002381	3.02E-05	5.34E-04

nominally significant SNVs are highlighted in yellow

Table 4

Allele Frequency and Association tests in Hispanics

Gene	ID	TARGETED SEQ AFFECTED CARRIERS	TARGETED SEQ AFFECTED CARRIERS FREQ	CTRLFREQ	FAMILIAL CASE FREQ	FAMILIAL CTRL FREQ	BETA	P	ExAC FREQ IN LATINO	FISHER TEST P VALUE
BIN1	rs138047593	6	0.01705	0.0084	0.0859	0.0641	2.03	1.27E-05	2.60E-03	5.85E-04
CD2AP	rs138727736	2	0.00568	0.0108	0.0238	0.0128	1.26	3.04E-02	3.78E-03	3.91E-01
CD2AP	rs116754410	2	0.00568	0.0160	0.0323	0.0128	0.69	3.37E-01	1.77E-03	1.41E-01
EPHA1	rs202178565	2	0.00568	0.0011	0.0078	0	3.44	1.25E-04	8.64E-05	2.55E-03
CLU	rs41276297	1	0.00284	0.0012	0	0			1.04E-03	3.23E-01
MS4A6A	rs138650483	1	0.00284	0.0049	0.0085	0	0.63	4.56E-01	4.16E-03	1.00E+00
PICALM	rs147556602	1	0.00472	0.0037	0.0154	0	1.21	1.81E-01	7.84E-04	1.67E-01

nominally significant SNVs are highlighted in yellow

Ann Neurol. Author manuscript; available in PMC 2016 September 01.

TABLE 5

Number of mutations (mutation rate per subject) in the three different mutation classes for each gene in NIA-LOAD, Toronto and HISPANIC Datasets

		CASES			CONTROLS		
		C1*	C2*	C3*	C1	C2	C3
NIA-LOAD DATASET	ABCA7	12 (0.100)	15 (0.125)	2 (0.0170)	0	0	0
	BIN1	1 (0.008)	2 (0.017)	3 (0.025)	0	0	2
	CD2AP	2 (0.017)	2 (0.017)	2 (0.017)	1	1	1
	CLU	1 (0.008)	1 (0.008)	1 (0.008)	0	0	0
	CR1	0	1 (0.008)	2 (0.017)	0	0	0
	EPHA1	3 (0.025)	3 (0.025)	2 (0.017)	0	0	0
	MS4A6A	1 (0.008)	1 (0.008)	0	1	1	0
	PICALM	0	3 (0.025)	0	0	0	0
	TOTAL	20	28	12	2	2	3
Toronto DATASET	ABCA7	16 (0.075)	16 (0.075)	5 (0.023)	1	1	0
	BIN1	8 (0.038)	8 (0.038)	4 (0.019)	0	0	4
	CD2AP	4 (0.019)	8 (0.038)	3 (0.014)	1	2	0
	CLU	0	2 (0.009)	5 (0.023)	0	0	1
	CR1	1 (0.005)	1 (0.005)	9 (0.042)	0	0	1
	EPHA1	1 (0.005)	5 (0.023)	6 (0.028)	0	0	2
	MS4A6A	1 (0.005)	2 (0.009)	0	0	0	0
	PICALM	2 (0.009)	2 (0.009)	1 (0.005)	1	1	0
	TOTAL	33	44	33	3	4	8
BURDEN TEST	N	53	72	45	5	6	11
	ENRICHMENT (P)	2.76 (0.02)	3.1 (0.002)	1.07 (0.59)			
HISPANIC DATASET	ABCA7	15 (0.085)	17 (0.097)	12 (0.068)			
	BIN1	2 (0.011)	9 (0.051)	8 (0.045)			
	CD2AP	0	4 (0.023)	0			
	CLU	1 (0.006)	4 (0.023)	5 (0.028)			
	CR1	0	4 (0.023)	2 (0.011)			
	EPHA1	8 (0.045)	8 (0.045)	6 (0.034)			

		CASES			CONTROLS		
		C1*	C2*	C3*	C1	C2	C3
	MS4A6A	0	1 (0.006)	0			
	PICALM	0	1 (0.006)	0			
	TOTAL	26	48	33			

*Class I: Loss of Function (stopgain, stoploss) and Damaging Missense Mutations

*Class II: All Non-Synonymous Mutations

*Class III: All Synonymous Mutations