Molecular Psychiatry (2016) 21, 916–924

www.nature.com/mp

OPEN

ORIGINAL ARTICLE

*APOE*E2* allele delays age of onset in *PSEN1* E280A Alzheimer's disease

JI Vélez[1,2,12], F Lopera[2,12], D Sepulveda-Falla[2,3,12], HR Patel[1], AS Johar[1], A Chuah[4], C Tobón[2], D Rivera[2], A Villegas[2], Y Cai[1], K Peng[5], R Arkell[6], FX Castellanos[7,8], SJ Andrews[9], MF Silva Lara[1], PK Creagh[1], S Easteal[9], J de Leon[10], ML Wong[11], J Licinio[11,12], CA Mastronardi[1,11,12] and M Arcos-Burgos[1,2,12]

Alzheimer's disease (AD) age of onset (ADAOO) varies greatly between individuals, with unique causal mutations suggesting the role of modifying genetic and environmental interactions. We analyzed ~50 000 common and rare functional genomic variants from 71 individuals of the 'Paisa' pedigree, the world's largest pedigree segregating a severe form of early-onset AD, who were affected carriers of the fully penetrant E280A mutation in the presenilin-1 (*PSEN1*) gene. Affected carriers with ages at the extremes of the ADAOO distribution (30s–70s age range), and linear mixed-effects models were used to build single-locus regression models outlining the ADAOO. We identified the rs7412 (*APOE*E2* allele) as a whole exome-wide ADAOO modifier that delays ADAOO by ~12 years ($\beta = 11.74$, 95% confidence interval (CI): 8.07–15.41, $P = 6.31 \times 10^{-8}$, $P_{FDR} = 2.48 \times 10^{-3}$). Subsequently, to evaluate comprehensively the APOE (apolipoprotein E) haplotype variants (*E1/E2/E3/E4*), the markers rs7412 and rs429358 were genotyped in 93 AD affected carriers of the E280A mutation. We found that the *APOE*E2* allele, and not *APOE*E4*, modifies ADAOO in carriers of the E280A mutation ($\beta = 8.24$, 95% CI: 4.45–12.01, $P = 3.84 \times 10^{-5}$). Exploratory linear mixed-effects multilocus analysis suggested that other functional variants harbored in genes involved in cell proliferation, protein degradation, apoptotic and immune dysregulation processes (i.e., *GPR20*, *TRIM22*, *FCRL5*, *AOAH*, *PINLYP*, *IFI16*, *RC3H1* and *DFNA5*) might interact with the *APOE*E2* allele. Interestingly, suggestive evidence as an ADAOO modifier was found for one of these variants (*GPR20*) in a set of patients with sporadic AD from the Paisa genetic isolate. This is the first study demonstrating that the *APOE*E2* allele modifies the natural history of AD typified by the age of onset in E280A mutation carriers. To the best of our knowledge, this is the largest analyzed sample of patients with a unique mutation sharing uniform environment. Formal replication of our results in other populations and in other forms of AD will be crucial for prediction, follow-up and presumably developing new therapeutic strategies for patients either at risk or affected by AD.

Molecular Psychiatry (2016) **21,** 916–924; doi:10.1038/mp.2015.177; published online 1 December 2015

INTRODUCTION

For the past three decades, we have studied the world's largest known pedigree in which a *presenilin-1* (*PSEN1*) mutation (p.Glu280Ala, E280A), often referred as the *Paisa* mutation, dominantly cosegregates with early-onset of Alzheimer's disease (AD).[1] This founder effect dates from the Spanish Conquistadors colonizing Colombia during the early sixteenth century.[1–4] Of the more than 5000 individuals descended from the original founder, we have enrolled 1784 in a comprehensive ongoing clinical monitoring study. Of 1181 genotyped participants, 459 are mutation carriers and 722 are non-carriers of the mutation.[4]

The importance of this pedigree is highlighted by the recent decision of the National Institutes of Health (NIH) to launch the first prevention trial for AD, with the Paisa *PSEN1* pedigree as one of the principal focuses. Along with the presence of exhaustive and detailed comprehensive medical records of thousands of individuals, this pedigree originated from a founder effect that makes it a valuable resource for genetic research,[2,5] and for the development of biomarkers for predicting and following up the natural history of AD.[6]

Although the median age of AD age of onset (ADAOO) in patients with the E280A *Paisa* mutation is 49 years, the ADAOO extends widely from the early 30s to the late 70s.[4] We considered that the high variability in ADAOO of this pedigree could include patients with an extreme phenotype (i.e., group of signs or symptoms departing from the disease's natural history that would manifest in patients with an extreme early or late ADAOO (e.g., 30s vs 70s, respectively),[7–9] supporting the hypothesis that the

[1]Genomics and Predictive Medicine Group, Department of Genome Sciences, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia; [2]Neuroscience Research Group, University of Antioquia, Medellín, Colombia; [3]Institute of Neuropathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [4]Genome Discovery Unit, Department of Genome Sciences, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia; [5]Biomolecular Resource Facility, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia; [6]Early Mammalian Development Laboratory, Research School of Biology, The Australian National University, Canberra, ACT, Australia; [7]NYU Child Study Center, NYU Langone Medical Center, New York, NY, USA; [8]Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, USA; [9]Genome Diversity and Health Group, Department of Genome Sciences, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia; [10]Mental Health Research Center at Eastern State Hospital, University of Kentucky, Lexington, KY, USA and [11]South Australian Health and Medical Research Institute and Department of Psychiatry, School of Medicine, Flinders University, Adelaide, SA, Australia. Correspondence: Dr M Arcos-Burgos, Genomics and Predictive Medicine Group, Department of Genome Sciences, John Curtin School of Medical Research, The Australian National University, Building 131, Garran Road, Canberra, ACT 2600, Australia.
E-mail: Mauricio.Arcos-Burgos@anu.edu.au
[12]These authors contributed equally to this work.
Received 20 April 2015; revised 7 October 2015; accepted 14 October 2015; published online 1 December 2015

variance in ADAOO is influenced by major modifiers. Since members of this pedigree share a relatively homogeneous environment and culture, we hypothesized that these modifiers of ADAOO could be genes shaping the natural history of cognitive decline. Indeed, a previous analysis by our group using a genome-wide association study approach found that genome-wide intronic variants significantly associated as ADAOO modifiers.[10] Other approaches using whole-genome sequencing have suggested some rare variants as potential modifiers of the ADAOO in this pedigree.[5]

To challenge the ADAOO variance, we scrutinized functional variants distributed through the whole exome in 71 *PSEN1* E280A mutation carriers, all of them descendants from the original *Paisa* pedigree founder. Subsequently, we evaluated some of these functional variants in 93 *PSEN1* E280A mutation carriers, and in a set of patients with sporadic AD (sAD) from the Paisa genetic isolate. Here, we disclose evidence that mutations harbored in *APOE* (apolipoprotein E), a gene implicated in the susceptibility and modification of AD risk, and additional new loci in *GPR20*, *TRIM22*, *FCRL5*, *AOAH*, *PINLYP*, *IFI16*, *RC3H1* and *DFNA5*, might modify the ADAOO and therefore substantially change the natural history of this condition. Furthermore, this oligogenic model exhibits substantial sensitivity and specificity to predict the ADAOO.

MATERIALS AND METHODS

Patients

E280A pedigree. Detailed clinical assessment and ascertainment procedures of this pedigree have been presented elsewhere.[4,11–13] Briefly, we have collected data from participants including clinical evaluations, family history, comprehensive neurological and neuropsychological examinations, functional MRI during face–name associative memory encoding and novel viewing and control tasks, and structural magnetic resonance imaging. Clinical, neurological and neuropsychological assessments at the Group of Neurosciences AD Clinic used a Spanish version of the CERAD (Consortium to Establish a Registry for Alzheimer's Disease) evaluation battery[14] adapted for the cultural and linguistic characteristics specific to this population[4,11–13] (described in detail in the Supplementary Material). Patients were defined as affected by mild cognitive impairment based on the Petersen's criteria and as AD if the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition) criteria were met.[15,16] The Ethics Committee of the University of Antioquia approved this study.

From the 459 E280A mutation carriers, we ascertained 71 patients from the extremes of the ADAOO distribution (44 women (62%) and 27 men (38%)) (Supplementary Figure 1a). The ADAOO (mean ± s.d.) was 47.8 ± 5.8 years in these patients (Supplementary Figure 1a). Mean ADAOO did not differ significantly by gender (female: 47.6 ± 6.1; male: 48.4 ± 5.5, $P = 0.55$) (Supplementary Figure 1b). A total of 43 patients (26 women (60%) and 17 men (40%)) had an ADAOO below 48 years.[10] As intended, mean ADAOO differed significantly between patients with ADAOO $\geqslant 48$ and < 48 years (44.2 ± 2.48 vs 53.5 ± 5.13, $P = 1.49 \times 10^{-10}$). In those individuals with available information ($n = 57$), years of education ranged from 0 to 19 years: four patients (7%) never attended school, 28 (49%) finished primary school (grades 1–5), 21 (37%) finished high school (grades 6–11, inclusive) and only 4 (7%) had tertiary education. Mean ADAOO did not differ significantly across education groups ($F_{3,53} = 2.721$, $P = 0.053$) (Supplementary Figures 1c and d).

Cohort of sporadic cases. We assembled an independent sample of 128 sporadic sAD cases recruited from the metropolitan area of Medellin, Antioquia, Colombia, to test whether variants associated with ADAOO in the E280A pedigree were also associated with ADAOO in this sample. Although these individuals do not carry the *PSEN1* E280A mutation, several population genetic analyses have shown that the community inhabiting this area has not been subject to microdifferentiation and shares the same genetic background and genealogy as the population with the E280A mutation.[2,3] Therefore, our rational is that given this similar genetic background, the genetic load modifying gene effects predisposing to AD is common. As in the E280A pedigree, neurological and neuropsychological assessments of these patients with sAD were conducted at the Group of Neurosciences AD Clinic using the modified CERAD evaluation battery (Supplementary Material).

Fifty-four patients, placed at the extremes of the ADAOO distribution (43 women (80%) and 11 men (20%)), were selected for this study from our collection of 128 patients with sAD (Supplementary Figure 2a). We chose them because the lower ADAOO in these patients had a similar distribution to that of patients with the E280A mutation. Similarly, the upper limit was defined to keep a similar distance between the average ADAOO and the upper limit of the E280A cohort. The average ADAOO was 63.26 ± 6.94 years (Supplementary Figure 2a) in the 54 patients, with no statistically significant differences by gender (females: 63.83 ± 6.29; males: 62.4 ± 6.06, $P = 0.52$) (Supplementary Figure 2b) or education group when including individuals with at least 1 year of education ($F_{2,46} = 2.61$, $P = 0.08$) (Supplementary Figure 2c). The number of years of education ranged from 0 to 18 years: 1 patient had no information, 1 (2%) never attended school, 22 (42%) completed primary school, 23 (43%) completed high school and 7 (13%) attended tertiary education (Supplementary Figure 2d).

Genotyping

Whole-exome genotyping. One hundred and eleven individuals (57 with AD from the E280A pedigree and 54 individuals with sAD) were whole-exome genotyped using Illumina HumanExome BeadChip-12v1_A. This SNP-chip covers putative functional exonic variants selected from over 12 000 individual exome and whole-genome sequences, and consists of ~250 000 markers representing diverse populations (including European, African, Chinese and Hispanic individuals), and a range of common conditions such as type 2 diabetes, cancer, metabolic and psychiatric disorders. In addition to pure exonic variation, the HumanExome BeadChip-12v1_A (Illumina, San Diego, CA, USA) chip covers single-nucleotide polymorphisms (SNPs) in splice sites, in stop variants, in promoter regions and genome-wide association study tag markers, among other potentially functional variation. Samples with calls below Illumina's expected 99% SNP call rates were excluded.

Whole-exome capture. We performed whole-exome capture on 14 individuals with AD from the E280A pedigree. Genomic DNA was extracted from peripheral blood from all patients and processed by the Australian Genome Facility (Melbourne, VIC, Australia). DNA libraries were constructed from 1 μg of genomic DNA using an Illumina TruSeq Genomic DNA Library Kit (Illumina, San Diego, CA, USA), and libraries were multiplexed with six samples pooled together (500 ng each). Exons were enriched from 3 μg of pooled library DNA using an Illumina TruSeq Exome Enrichment Kit (Illumina, San Diego, CA, USA), and ran on a 100 base pair paired-end run on an Illumina HiSeq 2000 sequencer (Illumina). A total of 201 071 genomic regions (sampled at ~50× coverage) were surveyed using the whole-exome capture platform.

Sequencing image data were processed in real time using Illumina's Real-Time Analysis Software and converted to FASTQ files using the CASAVA pipeline (Illumina). The entire workflow of data curation and analysis for variant calling was developed by the Genome Discovery Unit at The Australian National University, and consists of the following key components: (i) quality assessment; (ii) read alignment; (iii) local realignment around the known and novel indel regions to refine indel boundaries; (iv) recalibration of base qualities; (v) variant calling; and (vi) assigning quality scores to variants. The resulting FASTQ files were further processed for variant analysis using Golden Helix's SNP variation suite (SVS) 8.3.0 (Golden Helix, Bozeman, MT, USA).

Genetic, statistical and bioinformatics analyses

Quality control, filtering and classification of functional variants. After importing the genetic data to Golden Helix's SVS 8.3.0, a single genetic data file was constructed by merging common and uncommon exonic variants from both the whole-exome genotyping and whole-exome capture platforms. Genotypes for 71 individuals from the E280A pedigree were obtained and quality control subsequently performed using the following criteria: (i) deviations from Hardy–Weinberg equilibrium with *P*-values $< 0.05/m$ (where m is the number of markers included for analysis); (ii) a minimum genotype call rate of 90%; (iii) presence of two alleles (i.e., we excluded monoallelic markers, and markers that were present in more than two alleles). Markers not meeting any of these criteria were excluded from analyses. Genotype and allelic frequencies were estimated by maximum likelihood. Following previous recommendations,[17]

variants with a minor allele frequency (MAF) $\geqslant 0.01$ were classified as common and as rare otherwise.

Exonic variants with potential functional effect were determined using the functional prediction information available in the dbNSFP_NS_Functional_Predictions GRCh_37 annotation track.[18] This filter uses SIFT, PolyPhen-2, Mutation Taster, Gerp^{++} and PhyloP[19–21] and is fully implemented in the Golden Helix SVS 8.3.0 Variant Classification module. This module was also used to examine interactions between variants and gene transcripts to classify variants based on their potential effect on genes. Variants were classified according to their position in a gene transcript. In addition, variants in coding exons were further classified according to their effect on the gene's protein sequence.

Genome-wide association study analysis of common and rare variants. We studied the association of common exonic functional variants (CEFVs) to ADAOO using single- and multilocus additive, dominant and recessive linear mixed-effect models (LMEMs)[22] with up to 10 steps in the backward/forward optimization algorithm. The advantage of these models is the inclusion of both fixed (genotype markers, sex and years of education) and random effects (family or population structure), the later to account for potential inbreeding by including a kinship matrix (i.e., the identity-by-descent matrix, which in our case was estimated between all pairs of individuals using markers excluded from the final analysis after linkage disequilibrium pruning). A single-locus LMEM assumes that all loci have a small effect on the trait, whereas a multilocus LMEM assumes that several loci have a large effect on the trait.[22] Both types of models are implemented in SVS 8.3.0. The optimal model was selected using a comprehensive exploration of multiple criteria including the Extended Bayes Information Criteria, the Modified Bayes Information Criteria and the Multiple Posterior Probability of Association. After the estimation process using the forward/backward algorithm is finished, the coefficients $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m$ were extracted and a hypothesis test of the form $H_{0,i}$: $\beta_i = 0$ vs $H_{1,i}$: $\beta_i \neq 0$ was performed for the *i*th CEFVs to obtain the corresponding *P*-value ($i = 1,2,\ldots,m$). Thus, the collection P_1, $P_2,\ldots,P_m$ of *P*-values were subsequently corrected for multiple testing using the false discovery rate (FDR),[23] and a method based on extreme-values theory.[24] Because the tests of hypothesis being performed are of the same type, this correction is to be performed on the resulting *m* *P*-values only.[23] Using a type I error probability of 5%, any FDR-corrected *P*-value $\leqslant 3.62 \times 10^{-2}$ is considered statistically significant. This threshold was derived after correcting all *m* *P*-values using the p.adjust function in R.[25] Similarly, any raw *P*-value $\leqslant 1.26 \times 10^{-6}$ is statistically significant after Bonferroni's correction. For the exploratory analysis, only CEFVs located in genes modifying ADAOO in the E280A pedigree were included for association analysis in the cohort of sporadic cases using LMEMs.

For the analysis of rare exonic functional variants (REFVs), regression- and permutation-based kernel-based adaptive cluster methods were used.[26] Kernel-based adaptive cluster, implemented in SVS 8.3.0, catalogs rare variant data within each of a number of regions into multimarker genotypes, and, as variants are rare, only a relatively few different multimarker genotypes are found in any given region. A special test is subsequently applied to determine their association with the (case–control) phenotype, weighting each multimarker genotype by how often that genotype was expected to occur according to both the data and the null hypothesis that there is no association between that genotype and the case–control status of the sample.[26] Thus, genotypes with high sample risks are given higher weights that can potentially separate causal from non-causal genotypes. Further, a one-sided test was applied because of the weighting procedure and the *P*-values were estimated using 10 000 permutations. Individuals with ADAOO $\geqslant 48$ years were defined as cases and as controls otherwise. This cutoff value was selected based on previous studies of the ADAOO in the E280A pedigree.[4,10]

Genomic/clinical-based predictive framework with ARPA. We used advanced recursive partitioning approach (ARPA) to construct a predictive decision tree-based model of AD status (ADAOO < 48 and $\geqslant 48$ years) in our patients with *PSEN1* E280A AD using functional genetic variants and other clinical factors.[6,27,28] Gender, years of education and CEFVs identified as ADAOO modifiers were used as predictors. ARPA offers fast solutions to reveal hidden complex substructures and provides non-biased statistical analyses of high dimensional seemingly unrelated data, and is widely used in predictive analyses as it accounts for nonlinear hidden interactions better than alternative methods and is independent of the type of data and of the data distribution type.[29] ARPA was applied using the Classification and Regression Tree (CART), Random Forest (RF) and TreeNet modules implemented in the Salford Predictive Modeller software suite (Salford Systems, San Diego, CA, USA).

CART is a nonparametric approach whereby a series of recursive subdivisions separate the data by dichotomization.[30] The aim is to identify, at each partition step, the best predictive variable and its best corresponding splitting value while optimizing a splitting criterion. As a result, the data set is successfully split into increasingly homogeneous subgroups.[30] We used a battery of different statistical criteria as splitting rules (including the Gini index, Entropy and Twoing) to determine the splitting rule mostly decreasing the relative cost of the tree while increasing the prediction accuracy of target variable categories. The best split at each dichotomous node was chosen by either a measure of between-node dissimilarity or an iterative hypothesis testing of all possible splits to find the most homogeneous split (lowest impurity).[30] Similarly, we used a wide range of empirical *prior* probabilities to model numerous scenarios recreating the distribution of the targeted variable categories in the population.[30] Subsequently, each terminal node was assigned to a class outcome. To avoid overfitting in the CART predictive model, and to ensure that the final splits were well substantiated, tree pruning was applied. During this procedure, predictor variables that were close competitors (surrogate predictors with comparable overall classification error to the optimal predictors) were pruned to eliminate redundant commonalities among variables, thus the most parsimonious tree had the lowest misclassification rate for an individual not included in the original data.[30] The final CART predictive model was selected based on the performance measures presented in Supplementary Table 4.

RF was conjointly applied with a bagging strategy to identify exactly the most important set of variables predicting AD status.[31] Unlike CART, RF uses a limited number of variables to derive each node while creating hundreds to thousands of trees, and has proven to be immune to overfitting.[31] In the RF strategy, variables that appeared repeatedly in trees as predictors were identified, and the misclassification rate was recorded. Finally, TreeNet was used as a complement to CART and RF strategies because it reaches levels of accuracy that are usually not attainable by either of the other two.[32] This algorithm generates thousands of small decision trees built in a sequential error-correcting process that converge to an accurate model.[32] Cross-validation training with all data and then indirectly testing with all the data was performed to derive honest assessments of the derived models and have a better view of their performance on future unseen data (i.e., review the stability of results across multiple subsets).[30] To do so, the data were randomly divided, with replacement, into 10 separate partitions (folds). The used TreeNet (Boosting) decreases both the bias and the variance of the learning process, resulting in statistics that are immune to either inflation or overfitting. As described in the Results section, TreeNet corroborated the results of CART in both the learning and test data sets with slightly higher performance measures. This indicates a substantial predictive power of our ARPA-based CART predictive framework for identifying E280A mutation carriers with early and late ADAOO.

Pathway and network analyses. To identify key physiological pathways and networks involving genes harboring those variants disclosed by the common variant analysis, and to evaluate overrepresentation of common either ontogenetic or cellular processes, we performed network and pathway enrichment analyses with MetaCore® version 6.20 build 66481 (Thomson Reuters, New York, NY, USA). Genes with potential functional effect were examined with the 'Analyse Network', 'Process Networks', 'Shortest Paths' and 'Direct Interactions' algorithms. This collection of analyses provides a heuristic interpretation of maps and networks and rich ontologies for diseases based on the biological role of candidate genes. The presence of artifacts in statistical analyses (which can arise from genes in the database that may be in the same network but have no functional connection or interaction with any gene from our filtered list) was minimized by only including nodes with direct physical interactions between the encoded proteins in the database (known as the high trust set).

RESULTS

Quality control and genetic population structure

After quality control, assembling and filtering process, a total of 49 191 common and rare variants with potential functional effects remained for genetic analyses (Figure 1a). We estimated genetic stratification (population subdivision) using the F_{st} statistic of S

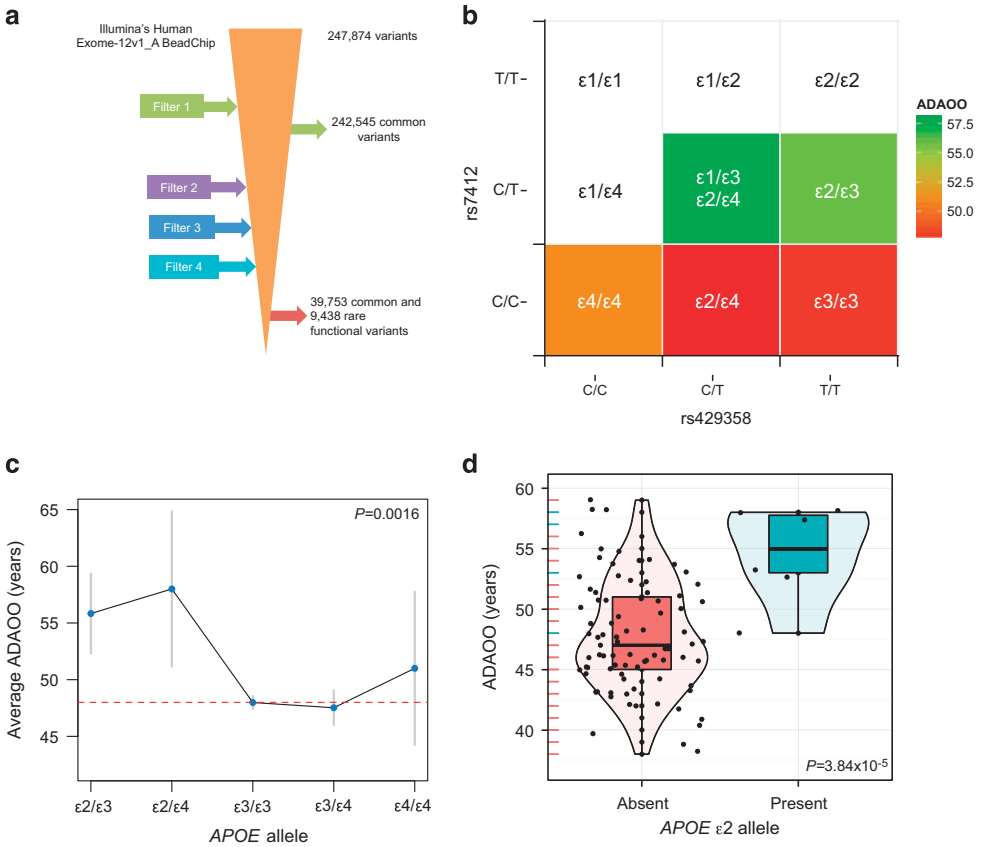


Figure 1. (**a**) Filtering process applied to exonic variants. *Filter 1* includes common and uncommon variants between the Illumina's HumanExome-12V1_A BeadChip and the whole-exome capture (see Patients and methods). *Filter 2* excludes variants with a genotype call rate $<90\%$, in Hardy–Weinberg disequilibrium and with one or more than two alleles. *Filter 3* excludes variants with minor allele frequency (MAF) $<1\%$ and *Filter 4* excludes nonfunctional variants. A total of 49 191 variants (39 753 common and 9438 rare) with potential functional effects remained for genetic analyses. (**b**) Average Alzheimer's disease (AD) age of onset (ADAOO) by *APOE* allele combination. (**c**) Average ADAOO (blue dot) as a function of the *APOE* alleles. The red and gray lines represent the ADAOO of 48 years and ± 1.5 s.d., respectively, the latter calculated using nonparametric bootstrap with $B=10\,000$ replicates. Analysis of variance (ANOVA) showed that the average ADAOO differs among allele groups ($F_{4,88}=4.74$, $P=0.00163$). (**d**) Effect of the presence/absence of the *APOE*E2* allele on ADAOO. A two-sample *t*-test indicates that the presence of this allele increases ADAOO by ~8.2 years (95% confidence interval (CI): 4.45–12.01, $P=3.84\times10^{-5}$) in presenilin-1 (*PSEN1*) E280A mutation carriers.

Wright. The estimated F_{st} value for these cohorts was of 0.0187 (there is formal consensus that F_{st} values >0.17 are correlated with microdifferentiation). Further, we estimated the kinship coefficient of relatedness. Pairs of fAD and sAD cases with F_{st} values >0.025 were discarded.

CEFVs modifying ADAOO

From the 39 753 CEFVs (Figure 1a), on average, ~2.83 CEFVs located within each gene (Supplementary Figure 4), we identified the rs7412, *APOE*E2* allele as a whole-exome-wide ADAOO modifier using a single-locus LMEM. This marker delays ADAOO by ~12 years ($\beta=11.74$, 95% CI: 8.07–15.41, $P=6.31\times10^{-8}$, $P_{FDR}=2.48\times10^{-3}$) (Table 1a).

Effect of *APOE*E2/E4* alleles on ADAOO

Given that rs7412 (*APOE*) is an ADAOO modifier in our patients with *PSEN1* E280A AD, we genotyped the rs429358 (*APOE*) marker to subsequently determine the effect of the *APOE* haplotype variants, defined by the *E1/E2/E3/E4* alleles (see Figure 1b), on ADAOO. A total of 93 individuals were genotyped. On average, the presence of the *APOE*E2* delayed the ADAOO by 8.24 years (95% CI: 4.45–12.01, $P=3.84\times10^{-5}$) when compared with its absence (Figures 1c and d).

A multilocus LMEM with nine steps in the backward/forward optimization algorithm was selected as the optimal model best explaining the ADAOO variance ($>90\%$ in total; Supplementary Figure 3a). Nine mutations harbored in *APOE* (rs7412, $P=5.44\times10^{-35}$), *GPR20* (rs36092215, $P=3.36\times10^{-26}$), *TRIM22* (rs12364019, $P=8.78\times10^{-19}$), *FCRL5* (rs16838748, $P=8.79\times10^{-14}$), *AOAH* (rs12701506, $P=7.26\times10^{-12}$), *PINLYP* (rs2682585, $P=2.55\times10^{-10}$), *IFI16* (rs62621173, $P=1.54\times10^{-9}$), *RC3H1* (rs10798302, $P=3.80\times10^{-8}$) and *DFNA5* (rs754554, $P=8.32\times10^{-6}$) were significantly associated as ADAOO modifiers (Table 1a and Supplementary Figure 3b and c). Variant rs12701506, located in the *AOAH* gene, is an intronic SNP anchored in a site encoding a strong enhancer (State-5) according to the chromatin state segmentation from ChiP-seq data. On the other hand, variant rs10798302 is located in an intergenic region close to the *RC3H1* gene, and is anchored in a CpG Island, DNaseI Hypersensitivity Uniform Peak according to the ENCODE/Analysis.

Exploratory analysis in a cohort of patients with sAD

From the 247 874 exonic variants available for genetic analysis in the sAD cohort, 17 variants were located within the genes significantly associated with ADAOO in the E280A cohort. A multilocus LMEM minimizing the Modified Bayes Information

Table 1a. Results of the association analysis for ADAOO in 71 patients with *PSEN1* E280A Alzheimer's disease

Chr	*SNP*[a]	*Position*	*Gene*	*Marker information*				*Single-locus linear mixed-effects model*		
				Ref/Alt	*MAF*	*CR*	*Change*	β (*s.e.*$_\beta$)	P-*value*	P_{FDR}
19	**Rs7412**	**45 412 079**	***APOE***	**C/T**	**0.044**	**1.000**	**p.Arg176Cys**	**11.74 (1.84)**	$\mathbf{6.31\times10^{-8}}$	$\mathbf{2.48\times10^{-3}}$
Chr	*SNP*[a]	*Position*	*Gene*	*Marker information*				*Multilocus linear mixed-effects model*		
				Ref/Alt	*MAF*	*CR*	*Change*	β (*s.e.*$_\beta$)	P-*value*	P_{FDR}
19	**Rs7412**	**45 412 079**	***APOE***	**C/T**	**0.044**	**1.000**	**p.Arg176Cys**	**17.45 (0.48)**	$\mathbf{5.44\times10^{-35}}$	$\mathbf{2.13\times10^{-30}}$
8	**Rs36092215**	**142 367 246**	***GPR20***	**G/A**	**0.036**	**0.982**	**p.Arg260Cys**	**12.12 (0.54)**	$\mathbf{3.36\times10^{-26}}$	$\mathbf{6.58\times10^{-22}}$
11	Rs12364019	5 730 343	*TRIM22*	G/A	0.018	1.000	p.Arg321Lys	−11.64 (0.79)	8.78×10^{-19}	1.15×10^{-14}
1	**Rs16838748**	**157 508 997**	***FCRL5***	**G/T**	**0.018**	**1.000**	**p.Asn427Lys**	**7.14 (0.68)**	$\mathbf{8.79\times10^{-14}}$	$\mathbf{8.61\times10^{-10}}$
7	Rs12701506	36 566 020	*AOAH*	G/A	0.096	1.000	[b]	−2.75 (0.30)	7.26×10^{-12}	5.69×10^{-8}
19	Rs2682585	44 081 288	*PINLYP*	A/G	0.219	1.000	p.His6Arg	−1.68 (0.21)	2.55×10^{-10}	1.67×10^{-6}
1	Rs62621173	159 021 506	*IFI16*	C/T	0.07	1.000	p.Ser512Phe	−2.80 (0.37)	1.54×10^{-9}	8.63×10^{-6}
1	**Rs10798302**	**173 987 798**	***RC3H1***[c]	**A/G**	**0.158**	**1.000**	[d]	**1.76 (0.27)**	$\mathbf{3.80\times10^{-8}}$	$\mathbf{1.86\times10^{-4}}$
7	Rs754554	24 758 818	*DFNA5*	G/T	0.132	1.000	p.Pro142Thr	−1.39 (0.28)	8.32×10^{-6}	3.62×10^{-2}

Abbreviations: β, regression coefficient; Chr, chromosome; CR, call rate; FDR, false discovery rate; MAF, minimum allele frequency; PSEN1, presenilin-1; Ref/Alt, reference/alternate allele; s.e.$_\beta$, standard error of β; SNP, single-nucleotide polymorphism. [a]UCSC GRCh37/hg19 coordinates. [b]Chromatin state segmentation strong enhancer state-5 from ChiP-seq data. [c]Nearest gene. [d]CpG islands, DNaseI hypersensitivity uniform peak from ENCODE/Analysis. No associations between genetic variants and sex were found (Supplementary Table 5). Bold variants decelerate AOO.

Table 1b. Findings in 54 patients with sporadic Alzheimer's disease

Chr	*SNP*[a]	*Position*	*Gene*	*Marker information*				*Multilocus linear mixed-effects model*[b]		
				Ref/Alt	*MAF*	*CR*	*Change*	β (*s.e.*$_\beta$)	P-value	P_{FDR}
8	Rs34591516	142 367 087	*GPR20*	C/T	0.037	1.000	p.Gly313Ser	−22.05 (6.87)	2.34×10^{-3}	4.44×10^{-2}

Abbreviations: AOO, age of onset; β, regression coefficient; Chr, chromosome; CR, call rate; FDR, false discovery rate; LMEM, linear mixed-effect model; MAF, minimum allele frequency; Ref/Alt, reference/alternate allele; s.e.$_\beta$, standard error of β; SNP, single-nucleotide polymorphism. [a]UCSC GRCh37/hg19 coordinates. [b]This modifier effect was subsequently confirmed using a single-locus LMEM ($\beta=-21.68$, s.e.$_\beta=6.96$, $P=3.1\times10$–3, $P_{FDR}=0.058$).

Criteria and Extended Bayes Information Criteria criteria and maximizing the Multiple Posterior Probability of Association criterion was selected. A CEFV harbored in *GPR20* (rs34591516, $P=2.34\times10^{-3}$) was found to modify ADAOO in our cohort of patients with sAD (Table 1b). This modifier effect was subsequently confirmed using a single-locus LMEM ($\beta=-21.68$, s.e.$_\beta=6.96$, $P=3.1\times10$–3, $P_{FDR}=0.058$).

Rare exonic functional variants modifying ADAOO

A total of 9438 functional rare variants were available after quality control, assembly and filtering process. Regression-based kernel-based adaptive cluster analysis disclosed nominal associations between transcripts in the *PDZD2* ($P=1.80\times10^{-2}$) and *ATM* ($P=2.4\times10^{-2}$) genes, as well as a borderline nominal association with transcripts in *C10orf12* ($P=5.0\times10^{-2}$) (Supplementary Table 1a). Permutation-based kernel-based adaptive cluster confirmed the associations in *PDZD2* ($P=1.40\times10^{-2}$) and *ATM* ($P=1.40\times10^{-2}$), and revealed a borderline nominal association between rare variants within *SDK2* ($P=5.29\times10^{-2}$) and ADAOO (Supplementary Table 1b). These results were corroborated by using optimized Sequence Kernel Association Test (Supplementary Table 1c).

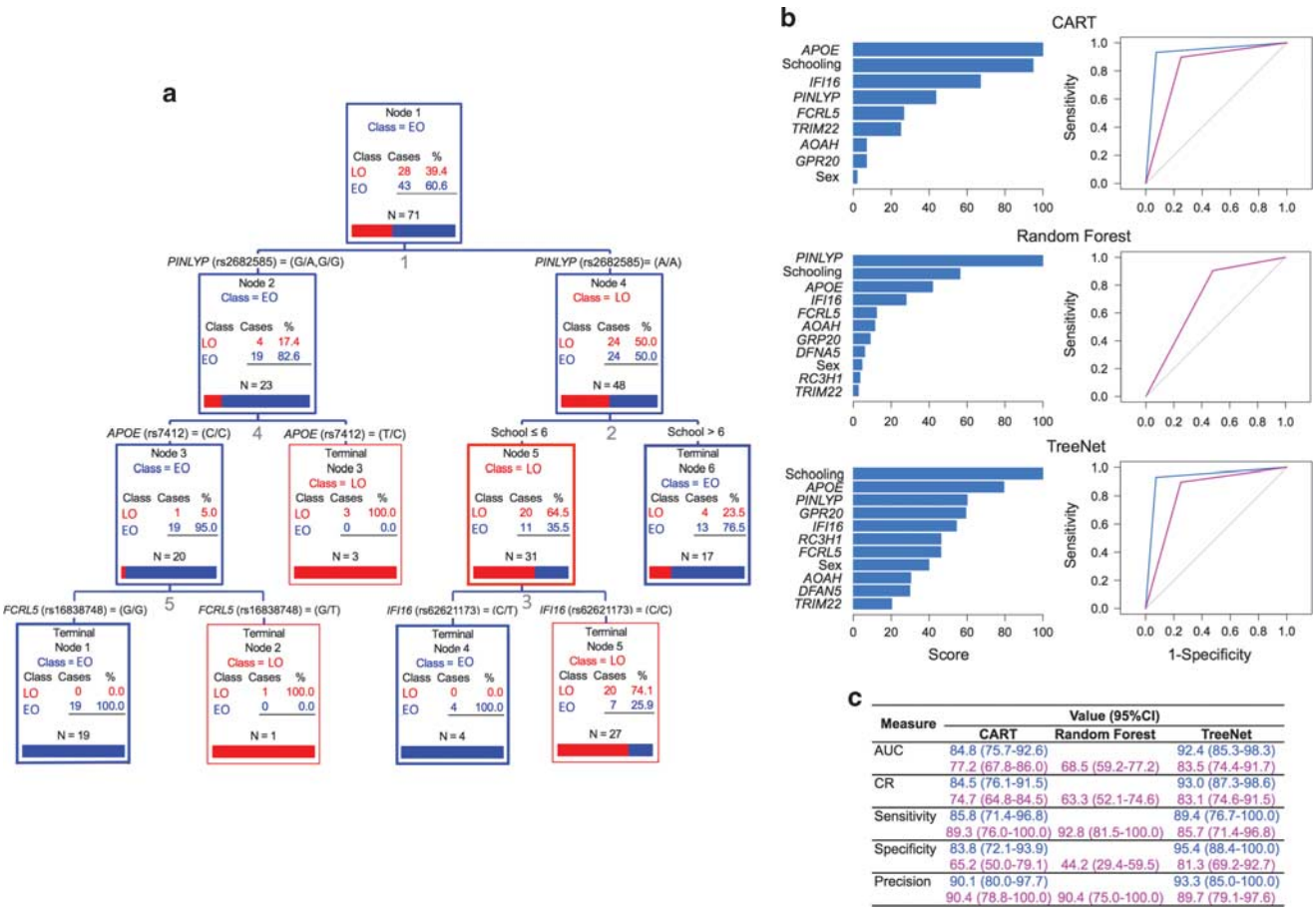
ARPA-based clinical diagnostic tool

A three-level tree with six terminal nodes was derived by CART, to identify E280A mutation carriers with late ADAOO (⩾48 years) and early ADAOO (<48 years). Validation of this predictive model via RF and TreeNet produces comparable results (see below). Splitting nodes involved years of education and variants rs2682585 (*PINLYP*), rs7412 (*APOE*), rs16838748 (*FCRL5*) and rs62621173 (*IFI16*) (Figure 2a).

The presence of two copies of the A allele in rs2682585 identifies 50% of the E280A carriers with late ADAOO (node 4, $n=24$). In the second split, E280A carriers with the A/A genotype in rs2682585 and six or fewer years of education were mostly identified as having late ADAOO (node 5, $n=20$, 64.5%), whereas attending 6 or more years of education classified them as having early ADAOO (node 6, $n=13$, 76.5%). In the third split, the presence of the C/C genotype in rs62621173 classified most of the E280A carriers with late ADAOO (terminal node 5, $n=20$, 74.1%), whereas the presence of the C/T genotype in rs62621173 classified all of the E280A carriers with early ADAOO (terminal node 4, $n=4$, 100%) (Figure 2a).

Subsequently, individuals having one or two copies of the G allele in rs2682585 were mostly identified as having early ADAOO (node 2, $n=23$, 82.6%) (Figure 2a, left). In split 4, these number of copies of the G allele in rs2682585 and the C/C genotype in rs7412 (*APOE*) correctly classified all E280A carriers with early ADAOO (node 3, $n=20$), whereas the T/C genotype discriminated 75% of those individuals with late ADAOO, suggesting a potential gene × gene interactions between *PINLYP* and *APOE* to modify the delaying effect of the *APOE*E2* allele on the ADAOO in *PSEN1* E280A mutation carriers (Figure 1d). Finally, split 5 identified the remaining individual with late ADAOO based on the G/T genotype in rs16838748 (*FCRL5*) (Figure 2a, bottom).

The variable importance and receiver operating characteristic (ROC) curves for the CART, RF and TreeNet strategies are shown in



c

Measure	Value (95%CI)		
	CART	Random Forest	TreeNet
AUC	84.8 (75.7-92.6)		92.4 (85.3-98.3)
	77.2 (67.8-86.0)	68.5 (59.2-77.2)	83.5 (74.4-91.7)
CR	84.5 (76.1-91.5)		93.0 (87.3-98.6)
	74.7 (64.8-84.5)	63.3 (52.1-74.6)	83.1 (74.6-91.5)
Sensitivity	85.8 (71.4-96.8)		89.4 (76.7-100.0)
	89.3 (76.0-100.0)	92.8 (81.5-100.0)	85.7 (71.4-96.8)
Specificity	83.8 (72.1-93.9)		95.4 (88.4-100.0)
	65.2 (50.0-79.1)	44.2 (29.4-59.5)	81.3 (69.2-92.7)
Precision	90.1 (80.0-97.7)		93.3 (85.0-100.0)
	90.4 (78.8-100.0)	90.4 (75.0-100.0)	89.7 (79.1-97.6)

Figure 2. (**a**) Classification tree for predicting late- (LO) and early-onset (EO) Alzheimer's disease in E280A mutation carriers. Numbers in gray represent the split number, and *N* the sample size within each node. (**b**) Variable importance (left) and receiver operating characteristic (ROC) curve (right) for the Classification and Regression Tree (CART), Random Forest and TreeNet strategies. (**c**) Performance measures for the learning (blue) and test (pink) data sets for each model (**b**, right panel). AUC, area under the curve; CI, confidence interval; CR, classification rate.

Figure 2b. Although similar results were obtained with all strategies, CART included fewer variables than RF and TreeNet. Overall, these results show that (i) the top 5 variables included in the final model are comparable among strategies; and (ii) years of education and the genotype in rs2682585 (*PINLYP*) and rs7412 (*APOE*) provide the most consistent set of variables for differentiating patients by ADAOO.

Figure 2c displays the performance measures for the testing and learning data sets (see Supplementary Tables 4a and b for more details). For the learning data set, CART estimated an area under the curve of 84.9 (95% CI = 75.7–92.6), classification rate of 84.5 (95% CI = 76.1–91.5), sensitivity of 85.8 (95% CI = 71.4–96.8), specificity of 83.8 (95% CI = 72.1–93.9) and precision of 90.1 (95% CI = 80.0–97.7), with overlapping 95% CI for the learning data set based on 10-fold cross-validation (Figure 2c). TreeNet corroborated these results in both the learning and test data sets with slightly higher performance measures than those produced by CART. On the other hand, RF in the testing data set produced similar point estimates for sensitivity and precision, and overlapping CIs to those from CART and TreeNet for other performance measures. Altogether, these measures indicate substantial predictive power of these genetic variants along with years of school attendance for identifying E280A mutation carriers with early and late ADAOO when combined in an ARPA-based CART predictive framework (Figure 2a).

We also have used the age of onset as an interval variable and prediction for the best fitting generalized boosting regression model was attempted. We found very similar patterns of variant inclusion and prediction when using the ADAOO as a continuous variable instead of one dichotomized into binary classes (see Supplementary Figure 7).

Pathway enrichment analysis

The pathway, network and enrichment analysis (using the 'Shortest Paths' algorithm) disclosed statistically significant involvement of *APOE, TRIM22, IFI16, RC3H1* and *DFAN* genes, interacting with *PSEN1* in important physiological pathways, and gene ontology (GO) processes of apoptosis and immunological response, as well as in diseases such as AD ($P = 3.34 \times 10^{-5}$), early-onset AD ($P = 1.68 \times 10^{-2}$), delirium, dementia, amnestic and cognitive disorders ($P = 3.13 \times 10^{-5}$), frontotemporal lobar dementia and degeneration ($P = 4.3 \times 10^{-4}$), and neurodegenerative diseases ($P = 5.85 \times 10^{-4}$) (Figure 3 and Supplementary Table 2).

DISCUSSION

Given the modest outcomes of the common disease-common allele hypothesis,[33] new genomic approaches are needed in complex genetic disorders. Recently, a comprehensive alternative approach has emerged, which assesses genetic risk in terms of nonlinear interactions among genetic variants of major effect (i.e., functional mutations).[27,34–38] For this approach, the identification of extreme phenotypes in patients ascertained either from

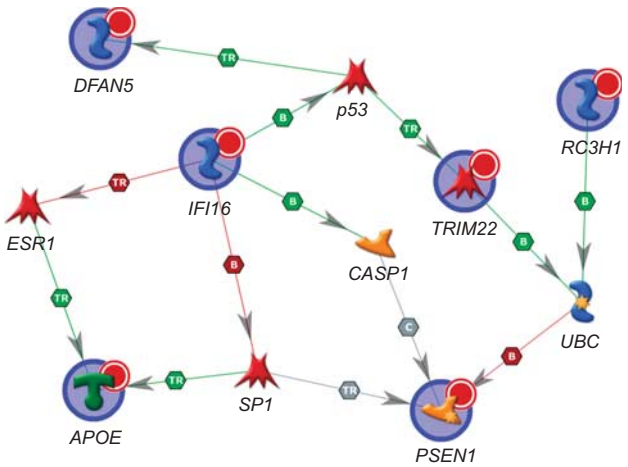


Figure 3. Resulting network involving genes harboring Alzheimer's disease age of onset (ADAOO) modifier mutations (red dot) in presenilin-1 (*PSEN1*) E280A Alzheimer's disease. Here, the 'Shortest Path' algorithm was used. B, binding; C, cleavage; gray, unspecified; green, positive/activation; red, negative/inhibition; TR, transcription/regulation.

extended and multigenerational pedigrees or homogeneous cohorts from genetic isolates is recommended.[2,10,39,40]

In this manuscript, we demonstrate the effectiveness of this strategy by identifying an oligogenic model comprised of functional variants of major effect harbored in the *APOE*, *GPR20*, *TRIM22*, *FCRL5*, *AOAH*, *PINLYP*, *IFI16*, *RC3H1* and *DFNA5* genes. Despite the relatively small sample size of our cohort, we have demonstrated the predictive efficiency of this model (Figure 2c) to delineate the ADAOO in E280A mutation carriers when clinical and demographic data are used in an ARPA-based classification tree. Thus, it is very unlikely that this tree might be the result of overfitting as using other techniques such as RF and TreeNet, which are immune to unbalanced group sizes, resulting in the same qualitative solution. Given the excellent performance of this tree in terms of sensitivity, specificity and precision (Figure 2c), we believe that this predictive framework can be used not only for predicting the ADAOO in the E280A pedigree but also for monitoring patients with AD at follow-up visits in future clinical trials. Having previously shown that imaging approaches such as ^{1}HMRI could predict the onset of symptoms in the same pedigree,[6] the next step would be to combine genomic, imaging and clinical data in an integrative predictive framework.

The characterization of *APOE* variants as one of the main drivers of the ADAOO model deserves further consideration. Indeed, preliminary analyses of this cohort were in apparent conflict with our current results. The first study using data from 31 E280A mutation carriers of the Paisa pedigree found nonsignificant effects of *APOE* variants on the ADAOO.[41] The second study expanded the number of patients with the E280A mutation to 52 and found that carriers of the *APOE* ε4 allele were more likely to develop AD at an earlier age than non-carriers (hazard ratio = 2.07; 95% CI = 1.07–3.99; $P < 0.03$), and that the *APOE*E2* allele had a modest statistically nonsignificant ADAOO decelerator effect.[42] By increasing the sample size to 93 patients with AD carrying the E280A mutation, we now show that individuals carrying the *APOE*E2* allele develop AD at a later age. We also found that individuals carrying the *APOE*E4* allele displayed a nonsignificant trend to develop AD at an early age. Our calculations show that this sample size is sufficient to detect a small-to-large effect size with >90% power for a type I error probability of 5% (Supplementary Figures 5 and 6).

The resulting network from the pathway and enrichment analysis shown in Figure 3 also includes the *ESR1*, *SP1*, *CASP1* and *UBC* genes, and the p53 tumor suppressor protein, all of which interact with some of our ADAOO modifier genes. In particular, p53 is activated by *IFI16*, and positively regulates both *DFAN5* and *TRIM22*. A mutation of either of these genes may affect this pathway and result in the upregulation of p53 in AD.[43]

CASP1, an essential component of NLPR3 inflammasomes, encodes a cysteine protease that is largely overexpressed in brains of patients with AD.[44] One of the key biological functions of CASP1 is the cleavage of prointerleukin-1β (IL-1β) into the mature biological active cytokine, which is largely overexpressed in brains of patients with AD.[44] Preclinical studies support the concept that CASP1-directed increase of IL-1β can augment the deposition of β-amyloid within the brain to worsen Alzheimer's symptoms.[44] Thus, it has been recently hypothesized that the NLRP3 inflammasome could be a potential target to treat AD.[44] Interestingly, IFI16, the product of the ADAOO accelerator gene *IFI16* identified in our study, binds and positively coactivates CASP1, which could trigger the inflammatory cascade to worsen Alzheimer's symptoms.[45] On the other hand, IFI16 interacts with and inhibits SP1 interaction to its cognate binding sites on DNA.[46] It is noteworthy that SP1, a transcription factor that regulates the expression of several amyloid and tau-related genes,[47] has been found to be abnormally expressed in the frontal cortex and hippocampus of patients with AD.[48] Hence, mutations in *IFI16* could result in the exacerbation of the activation of the inflammatory cascade orchestrated by CASP1, and/or the increased transcriptional activity of SP1 controlling the expression of key genes that are involved in the pathogenesis of AD. Our pathway and enrichment analysis also suggests that *IFI16* activates *ESR1*, which in turn leads to the positive regulation of *APOE* whilst *SP1* positively regulates *APOE* and *PSEN1* (although these mechanisms are not yet well understood). In search for therapeutic targets, our data support that *ESR1*, *SP1* and *CASP1* could be important alternatives. Additional relevant GO processes involving at least two of the ADAOO modifier genes reported here are presented in Supplementary Table 3.

TRIM22 is an E3 ubiquitin ligase that is a member of the tripartite motif (TRIM) family, which includes three zinc-binding domains, a RING finger, a B-box type 1 and a B-box type 2, and a coiled-coil region. Although there are no reports relating TRIM22 with AD, TRIM11, another member of this family, was shown to bind to and destabilize humanin, a neuroprotective peptide that suppresses AD-related neurotoxicity.[49] Interestingly, *RC3H1* also encodes a protein that has an amino-terminal RING-1 zinc-finger, which is characteristic to that of the E3 ubiquitin ligase family.[50] In the pathway enrichment analysis (Figure 3), it is predicted that both TRIM22 and RC3H1 bind to UBC, which in turn bind to PSEN1 to cause inhibitory effects. In previous studies it has been shown that presenilin proteins can be ubiquitinated.[51,52] Moreover, there is evidence suggesting that presenilin ubiquitination can cause reduction of endoproteolisis, which in turn reduces the formation of two fragments (presenilin N- and C-terminal fragments) that are essential for the γ-secretase activity.[51] Thus, mutations in these genes may affect PSEN1 ubiquitination, thereby potentially leading to protein degradation and favoring β-amyloid and hyperphosphorylated Tau deposition.

GPR20 was not modeled in the pathway enrichment analysis shown in Figure 3. *GPR20* belongs to the family of the G-protein-coupled receptors, which participate in intracellular second messenger systems triggered by different hormones and neurotransmitters, and amyloid β.[53–55] *GPR20* is expressed in AD brain.[56]

Three novel variants modifying the ADAOO are located in *FCRL5*, *AOAH* and *PINLYP*. Despite not being modeled in the pathway enrichment analysis and the lack of reported evidence of their possible relationship with AD, these genes encode key products having relevant functions within the immune system,

which seems to influence several neurodegenerative and mental diseases.[57,58] *FCRL5* encodes Fc receptor-like 5 that regulates B-cell antigen receptor signaling.[59] Even though the expression of FCRL5 within the brain has not yet been reported, other Fc receptors were found to be expressed in microglia, the brain-resident macrophages.[60] *AOAH* encodes acyloxycyl hydrolase, which is an enzyme expressed in antigen-presenting cells that deacylates and inactivates endotoxin.[61] Finally, *PINLYP* encodes phospholipase A2 inhibitor and LY6/PLAUR domain containing. Although there is no much functional evidence reported about this gene product, it is well known that phospholipase A2 releases arachidonic acid, which is a substrate used for the synthesis of potent proinflammatory factors such as prostaglandins. Thus, it can be hypothesized that phospholipase A2 inhibition caused by the *PINLYP* product could favor an anti-inflammatory effect.

Limitations of this study include the fact that the oligogenic model was derived from a very unique form of *f*AD, and therefore it might only be applicable to carriers of the *PSEN1* mutation. However, the identification of one of these associated variants in patients with sAD shows a common gene effect predisposing to AD to that in the E280A pedigree, and suggests that these variants could be real major players modifying the natural history of the illness. The analysis of other AD cohorts from around the world would be the next focus of our research.

In summary, we have defined major mutations modifying ADAOO in members of a multigenerational extended family carrying the *PSEN1* E280A mutation. One of the modifier genes reported herein was also associated with the ADAOO in sporadic cases from the general population. This suggests that the ADAOO modifier effect is not unique to the Paisa pedigree; it may be a general finding applicable to other forms of AD. Of major importance is the highlighting of the *APOE*E2* allele as an ADAOO decelerator. This finding is consistent with recent research,[62] and suggests that *PSEN1* and *APOE* may interact to modify ADAOO in these patients. Furthermore, using a subset of these variants, we constructed an accurate predictive framework to characterize AD patients in terms of early or late onset that can be used as a diagnostic tool during clinical assessment. Finally, the pathway enrichment analysis suggests that cell proliferation, protein degradation, apoptotic and dysregulation of immune processes are implicated in the variable onset of AD.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We express our highest appreciation to the patients and relatives enrolled in this study for more than 20 years. This study was financed by a research grant from the Australian National University (ANU) (to MA-B) to launch his laboratory, and COLCIENCIAS and the University of Antioquia, Grant 1115-408-20543. JIV was supported by the Eccles Scholarship in Medical Sciences, the Fenner Merit Scholarship and the ANU High Degree Research scholarships. JIV, ASJ and SJA are doctoral students at the ANU. Some of this work is to be presented in partial fulfillment of PhD degree requirements. DS-F was supported by the Hamburg State Ministry of Science and Research, Landesforschungsförderung 'Molekulare mechanismen der netzwerkmodifizierung'. The sponsor of the study has no role in the study design, data collection, data analysis, data interpretation or writing of the reports. JIV, FL, DS-F and MA-B have full access to all the data in the study. JIV, CAM and MA-B are responsible for submitting this work for publication.

REFERENCES

1 Lopera F, Ardilla A, Martinez A, Madrigal L, Arango-Viana JC, Lemere CA *et al.* Clinical features of early-onset Alzheimer disease in a large kindred with an E280A presenilin-1 mutation. *JAMA* 1997; **277**: 793–799.
2 Arcos-Burgos M, Muenke M. Genetics of population isolates. *Clin Genet* 2002; **61**: 233–247.
3 Bravo ML, Valenzuela CY, Arcos-Burgos OM. Polymorphisms and phyletic relationships of the Paisa community from Antioquia (Colombia). *Gene Geogr* 1996; **10**: 11–17.
4 Acosta-Baena N, Sepulveda-Falla D, Lopera-Gomez CM, Jaramillo-Elorza MC, Moreno S, Aguirre-Acevedo DC *et al.* Pre-dementia clinical stages in presenilin 1 E280A familial early-onset Alzheimer's disease: a retrospective cohort study. *Lancet Neurol* 2011; **10**: 213–220.
5 Lalli MA, Garcia G, Madrigal L, Arcos-Burgos M, Arcila ML, Kosik KS *et al.* Exploratory data from complete genomes of familial alzheimer disease age-at-onset outliers. *Hum Mutat* 2012; **33**: 1630–1634.
6 Londono AC, Castellanos FX, Arbelaez A, Ruiz A, Aguirre-Acevedo DC, Richardson AM *et al.* An 1H-MRS framework predicts the onset of Alzheimer's disease symptoms in PSEN1 mutation carriers. *Alzheimer's Dement* 2014; **10**: 552–561.
7 Barnett IJ, Lee S, Lin X. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genet Epidemiol* 2013; **37**: 142–151.
8 Johar AS, Anaya JM, Andrews D, Patel HR, Field M, Goodnow C *et al.* Candidate gene discovery in autoimmunity by using extreme phenotypes, next generation sequencing and whole exome capture. *Autoimmun Rev* 2014; **14**: 204–209.
9 Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genet Epidemiol* 2011; **35**: 790–799.
10 Velez JI, Chandrasekharappa SC, Henao E, Martinez AF, Harper U, Jones M *et al.* Pooling/bootstrap-based GWAS (pbGWAS) identifies new loci modifying the age of onset in PSEN1 p.Glu280Ala Alzheimer's disease. *Mol Psychiatry* 2013; **18**: 568–575.
11 Fleisher AS, Chen K, Quiroz YT, Jakimovich LJ, Gomez MG, Langois CM *et al.* Florbetapir PET analysis of amyloid-beta deposition in the presenilin 1 E280A autosomal dominant Alzheimer's disease kindred: a cross-sectional study. *Lancet Neurol* 2012; **11**: 1057–1065.
12 Reiman EM, Quiroz YT, Fleisher AS, Chen K, Velez-Pardo C, Jimenez-Del-Rio M *et al.* Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant Alzheimer's disease in the presenilin 1 E280A kindred: a case–control study. *Lancet Neurol* 2012; **11**: 1048–1056.
13 Reiman EM, Langbaum JB, Fleisher AS, Caselli RJ, Chen K, Ayutyanont N *et al.* Alzheimer's prevention initiative: a plan to accelerate the evaluation of presymptomatic treatments. *J Alzheimers Dis* 2011; **26**: 321–329.
14 Morris JC, Heyman A, Mohs RC, Hughes JP, van Belle G, Fillenbaum G *et al.* The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* 1989; **39**: 1159–1165.
15 Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E. Mild cognitive impairment: clinical characterization and outcome. *Arch Neurol* 1999; **56**: 303–308.
16 American Psychiatric Association. Diagnostic and statistical manual of mental disorders, 4th (edn). American Psychiatric Association: Washington, DC, 2000.
17 Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 2010; **11**: 773–785.
18 Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010; **6**: e1001025.
19 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.
20 Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003; **31**: 3812–3814.
21 Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010; **7**: 575–576.
22 Segura V, Vilhjalmsson BJ, Platt A, Korte A, Seren U, Long Q *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* 2012; **44**: 825–830.
23 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995; **57**: 289–300.
24 Vélez JI, Correa JC, Arcos-Burgos M. A new method for detecting significant *p*-values with applications to genetic data. *Rev Colomb Estad* 2014; **37**: 67–76.
25 R Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria, 2015, Available at: http://www.R-project.org.
26 Liu DJ, Leal SM. Replication strategies for rare variant complex trait association studies via next-generation sequencing. *Am J Hum Genet* 2010; **87**: 790–801.
27 Wong ML, Dong C, Andreev V, Arcos-Burgos M, Licinio J. Prediction of susceptibility to major depression by a model of interactions of multiple functional genetic variants and environmental factors. *Mol Psychiatry* 2012; **17**: 624–633.
28 Wong ML, Dong C, Flores DL, Ehrhart-Bornstein M, Bornstein S, Arcos-Burgos M *et al.* Clinical outcomes and genome-wide association for a brain methylation site

in an antidepressant pharmacogenetics study in Mexican Americans. *Am J Psychiatry* 2014; **171**: 1297–1309.
29 Rao DC. CAT scans, PET scans, and genomic scans. *Genet Epidemiol* 1998; **15**: 1–18.
30 Breiman L, Friedman JH, Olshen RA, Stone CH. Classification and regression trees. Wadsworth International Group: Belmont, CA, USA, 1984.
31 Breiman L. Random forests. In: Schapire RE (ed). Machine Learning, Vol 45. Manufactured in the Netherlands: Statistics Department, University of California. Kluwer Academic Publishers: Berkeley, CA, USA, 2001, pp 5–32.
32 Friedman JH. Greedy function approximation: A gradient boosting machine. Department of Statistics, University of Stanford: Stanfors, CA, USA, 1999.
33 Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010; **11**: 415–425.
34 Fearnhead NS, Wilding JL, Winney B, Tonks S, Bartlett S, Bicknell DC *et al.* Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc Natl Acad Sci USA* 2004; **101**: 15992–15997.
35 Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K *et al.* A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput Biol* 2010; **6**: e1000954.
36 Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet* 2010; **6**: e1001156.
37 Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008; **40**: 695–701.
38 Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* 2012; **109**: 1193–1198.
39 Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, Bjornsson S *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* 2012; **488**: 96–99.
40 Arcos-Burgos M, Jain M, Acosta MT, Shively S, Stanescu H, Wallis D *et al.* A common variant of the latrophilin 3 gene, LPHN3, confers susceptibility to ADHD and predicts effectiveness of stimulant medication. *Mol Psychiatry* 2010; **15**: 1053–1066.
41 Lendon CL, Martinez A, Behrens IM, Kosik KS, Madrigal L, Norton J *et al.* E280A PS-1 mutation causes Alzheimer's disease but age of onset is not modified by ApoE alleles. *Hum Mutat* 1997; **10**: 186–195.
42 Pastor P, Roe CM, Villegas A, Bedoya G, Chakraverty S, Garcia G *et al.* Apolipoprotein Eepsilon4 modifies Alzheimer's disease onset in an E280A PS1 kindred. *Ann Neurol* 2003; **54**: 163–169.
43 Hooper C, Meimaridou E, Tavassoli M, Melino G, Lovestone S, Killick R. p53 is upregulated in Alzheimer's disease and induces tau phosphorylation in HEK293a cells. *Neurosci Lett* 2007; **418**: 34–37.
44 Heneka MT, Kummer MP, Stutz A, Delekate A, Schwartz S, Vieira-Saecker A *et al.* NLRP3 is activated in Alzheimer's disease and contributes to pathology in APP/PS1 mice. *Nature* 2013; **493**: 674–678.
45 Monroe KM, Yang Z, Johnson JR, Geng X, Doitsh G, Krogan NJ *et al.* IFI16 DNA sensor is required for death of lymphoid CD4 T cells abortively infected with HIV. *Science* 2014; **343**: 428–432.
46 Gariano GR, Dell'Oste V, Bronzini M, Gatti D, Luganini A, De Andrea M *et al.* The intracellular DNA sensor IFI16 gene acts as restriction factor for human cytomegalovirus replication. *PLoS Pathog* 2012; **8**: e1002498.
47 Citron BA, Dennis JS, Zeitlin RS, Echeverria V. Transcription factor Sp1 dysregulation in Alzheimer's disease. *J Neurosci Res* 2008; **86**: 2499–2504.
48 Santpere G, Nieto M, Puig B, Ferrer I. Abnormal Sp1 transcription factor expression in Alzheimer disease and tauopathies. *Neurosci Lett* 2006; **397**: 30–34.
49 Niikura T, Hashimoto Y, Tajima H, Ishizaka M, Yamagishi Y, Kawasumi M *et al.* A tripartite motif protein TRIM11 binds and destabilizes Humanin, a neuroprotective peptide against Alzheimer's disease-relevant insults. *Eur J Neurosci* 2003; **17**: 1150–1158.
50 Vinuesa CG, Cook MC, Angelucci C, Athanasopoulos V, Rui L, Hill KM *et al.* A RING-type ubiquitin ligase family member required to repress follicular helper T cells and autoimmunity. *Nature* 2005; **435**: 452–458.
51 Massey LK, Mah AL, Monteiro MJ. Ubiquilin regulates presenilin endoproteolysis and modulates gamma-secretase components, Pen-2 and nicastrin. *Biochem J* 2005; **391**: 513–525.
52 Duggan SP, Yan R, McCarthy JV. A ubiquitin-binding CUE domain in presenilin-1 enables interaction with K63-linked polyubiquitin chains. *FEBS Lett* 2015; **589**: 1001–1008.
53 Ghanemi A. Targeting G protein coupled receptor-related pathways as emerging molecular therapies. *Saudi Pharm J* 2013; **23**: 115–129.
54 Rosenbaum DM, Rasmussen SG, Kobilka BK. The structure and function of G-protein-coupled receptors. *Nature* 2009; **459**: 356–363.
55 Thathiah A, De Strooper B. The role of G protein-coupled receptors in the pathology of Alzheimer's disease. *Nat Rev Neuroscience* 2011; **12**: 73–87.
56 Brueggemeier U, Geerts A, Golz S, Summer H. Diagnostics and therapeutics for diseases associated with g protein-coupled receptor 20 (gpr20). Google Patents, 2005.
57 McGeer PL, McGeer EG. The inflammatory response system of brain: Implications for therapy of Alzheimer and other neurodegenerative diseases. *Brain Res Rev* 1995; **21**: 195–218.
58 Block ML, Hong JS. Microglia and inflammation-mediated neurodegeneration: Multiple triggers with a common mechanism. *Progr Neurobiol* 2005; **76**: 77–98.
59 Franco A, Damdinsuren B, Ise T, Dement-Brown J, Li H, Nagata S *et al.* Human Fc receptor-like 5 binds intact IgG via mechanisms distinct from those of Fc receptors. *J Immunol* 2013; **190**: 5739–5746.
60 Lunnon K, Teeling JL, Tutt AL, Cragg MS, Glennie MJ, Perry VH. Systemic inflammation modulates Fc receptor expression on microglia during chronic neurodegeneration. *J Immunol* 2011; **186**: 7215–7224.
61 Janelsins BM, Lu M, Datta SK. Altered inactivation of commensal LPS due to acyloxyacyl hydrolase deficiency in colonic dendritic cells impairs mucosal Th17 immunity. *Proc Natl Acad Sci USA* 2014; **111**: 373–378.
62 Suri S, Heise V, Trachtenberg AJ, Mackay CE. The forgotten APOE allele: a review of the evidence and suggested mechanisms for the protective effect of APOE varepsilon2. *Neurosci Biobehav Rev* 2013; **37**: 2878–2886.

Supplementary Information accompanies the paper on the Molecular Psychiatry website (http://www.nature.com/mp)