Received: 30 August 2021 | Revised: 8 February 2022 | Accepted: 10 February 2022 | Published online: 8 April 2022

DOI: 10.1002/dad2.12300

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

RESEARCH ARTICLE

Machine learning methods applied to genotyping data capture interactions between single nucleotide variants in late onset Alzheimer's disease

Magdalena Arnal Segura[1,2,3] | Giorgio Bini[2] | Dietmar Fernandez Orth[3] | Eleftherios Samaras[4] | Maya Kassis[4] | Fotis Aisopos[5] | Jordi Rambla De Argila[3] | George Paliouras[5] | Peter Garrard[4] | Claudia Giambartolomei[2] | Gian Gaetano Tartaglia[1,2,3,6]

[1] Department of Biology "Charles Darwin", Sapienza University of Rome, Rome, Italy

[2] Centre for Human Technologies, Istituto Italiano di Tecnologia (IIT), Genova, Italy

[3] Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona, Spain

[4] Stroke and Dementia Research Centre, St George's, University of London, London, UK

[5] Institute of Informatics and Telecommunications, NCSR Demokritos, Athens, Greece

[6] Catalan Institution for Research and Advanced Studies, ICREA, Barcelona, Spain

Correspondence
Gian Gaetano Tartaglia, Centre for Human Technologies, Istituto Italiano di Tecnologia, Via Enrico Melen, 83, 16152 Genova, Italy.
Email: gian.tartaglia@iit.it

Senior authors Claudia Giambartolomei & Gian Gaetano Tartaglia contributed equally to this study.

Funding information
H2020 project IASIS, Grant/Award Number: 727658; European Research Council ASTRA, Grant/Award Number: 855923; European Genome-phenome Archive (EGA); Marie Skłodowska-Curie grant agreement, Grant/Award Number: 754490–MINDED

Abstract

Introduction: Genome-wide association studies (GWAS) in late onset Alzheimer's disease (LOAD) provide lists of individual genetic determinants. However, GWAS do not capture the synergistic effects among multiple genetic variants and lack good specificity.

Methods: We applied tree-based machine learning algorithms (MLs) to discriminate LOAD (>700 individuals) and age-matched unaffected subjects in UK Biobank with single nucleotide variants (SNVs) from Alzheimer's disease (AD) studies, obtaining specific genomic profiles with the prioritized SNVs.

Results: MLs prioritized a set of SNVs located in genes *PVRL2*, *TOMM40*, *APOE*, and *APOC1*, also influencing gene expression and splicing. The genomic profiles in this region showed interaction patterns involving rs405509 and rs1160985, also present in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. rs405509 located in *APOE* promoter interacts with rs429358 among others, seemingly neutralizing their predisposing effect.

Discussion: Our approach efficiently discriminates LOAD from controls, capturing genomic profiles defined by interactions among SNVs in a hot-spot region.

KEYWORDS
Apolipoprotein E, genetic determinants, genomic interactions, genomic profiles, late onset Alzheimer's disease, machine learning, single nucleotide variants, variant prioritization

1 | INTRODUCTION

1.1 | Background

Alzheimer's disease (AD) is a neurodegenerative pathology and the most common cause of late life dementia with symptoms such as memory loss, language deficits, disorientation, mood changes, and in advanced stages with loss of vegetative function and eventually death.[1] Approximately 5% of the total number diagnosed with AD develop symptoms of dementia between the ages 45 and 65, and are designated as early onset Alzheimer's disease (EOAD).[2] Conversely, the prevalence of the disease in the population aged above 65 currently represent ≈95% of the total AD cases, and are designated as late onset Alzheimer disease (LOAD).[3]

The genomic characterization of AD has improved in the last decades thanks to the emergence of genome-wide association studies (GWAS).[4] However, these tools miss the synergistic effects caused by various genomic loci and lack good specificity due to the multiple testing problem and linkage disequilibrium (LD).[5] In this context, the selection of genetic determinants for the follow-up in laboratory and clinical studies remains a challenge, and most of the mechanisms in which the discovered predisposing and protective genetic alterations contribute to AD are still unknown.[6] In the case of LOAD, heritability is estimated to be ≈56% to 79%,[7] and Apolipoprotein E polymorphic alleles (APOE $\varepsilon 2/\varepsilon 3/\varepsilon 4$) are the major genetic determinants of susceptibility discovered until now.[8,9] Nevertheless, there are more candidate genes such as *TOMM40*, *PVRL2*, *ABCA7*, *ADAM10*, *BIN1*, *CLU*, and *CR1*, among others,[4,10] and nowadays, there is a growing consensus considering LOAD a polygenic risk disease.[11]

Machine learning (ML) methods are growing in popularity for their contributions to a wide range of fields including medicine.[12,13] ML classifiers have been previously implemented to classify AD using genotyping data reaching an accuracy of 0.84.[14,15] Additionally, they have been used in the post-GWAS prioritization of genomic variants in several diseases.[16,17,18] As ML algorithms work better with a limited set of predictors to be efficient, and the full set of single nucleotide variants (SNVs) in genotyping arrays is too large to reach a reasonable computational performance, a set of AD-related SNVs are typically preselected and used as predictors in the ML models. As input variables, ML approaches can accept a list of SNVs without any prior assumptions about the genetic contribution to the traits and the method itself calculates the importance of the SNVs during the learning step.

In this study, our initial aim was to classify individuals with LOAD and controls without any neurodegenerative disease both from the UK Biobank (UKB),[19] using ML methods and data from genotyping arrays. Our second aim was to select the SNVs with higher feature importance (FI) and retrieve a set of genomic profiles that are related to AD. We did a first selection of genomic variants considering previously reported SNVs related to AD in the DisGeNet[20] database. DisGeNet integrates data from curated resources such as ClinVar,[21] the GWAS Catalog,[22] and GWASdb.[23] As for the classification method, we tested three tree-based ML approaches, gradient boosted decision trees (GB), extremely randomized trees (ET), and random forest (RF). Tree-based algorithms perform yes/no decisions in branches leading to a sample's classification, which is particularly appropriate with categorical predictors such as SNVs. Here we show the utility of tree-based ML methods to classify LOAD, to prioritize a small set of SNVs related to the disease, and to draw distinct LOAD genomic profiles based on the interactions between these SNVs.

RESEARCH IN CONTEXT

1. **Systematic review**: We used a set of single nucleotide variants (SNVs) related to Alzheimer's disease (AD) and machine learning (ML) approaches to classify people with late onset Alzheimer's disease (LOAD) and controls from the UK Biobank (UKB), reaching an area under the receiver operating characteristic curve (AUC-ROC) from 0.80 to 0.90. The correctly classified genomic profiles built with the prioritized SNVs in ML were interpreted.
2. **Interpretation**: The genomic profiles obtained with UKB samples showed interaction patterns involving two SNVs, rs405509 and rs1160985, that were also present in the Azheimer's Disease Neuroimaging Initiative dataset (ADNI). ML approaches revealed an interaction between rs405509 located in apolipoprotein E promoter and the upstream SNVs rs429358, rs769449, and rs4420638 seemingly neutralizing their predisposing effect to the disease. These interactions in a hot-spot region of chromosome 19 are supported by the presence of expression and splicing quantitative trait locis.
3. **Future directions**: We propose our approach to: (a) classify individuals with LOAD; (b) provide genomic profiles linked to the disease, revealing synergistic effects between SNVs located in close proximity.

2 | METHODS

2.1 | Sample selection and clinical information in UK Biobank and Alzheimer's Disease Neuroimaging Initiative

From UKB,[19] a total number of 738 participants with AD and >70 years old, were selected using the International Classification of Disease, 10th revision (ICD-10) codes representative of AD, excluding the EOAD ICD-10 codes from the available hospitalization records (Table S1 in supporting information, page 1). In addition, 75,000 participants were selected as controls >70 years old and without any reported mental and behavioral disorder (ICD-10: F00–F99) or disease of the nervous system (ICD-10: G00–G99) in hospitalization records. Participants that requested to withdraw from UKB were excluded. Gender and age distribution of selected samples across conditions is shown in Figure S1A, B in supporting information. There were 327 samples in Alzheimer's Disease Neuroimaging Initiative 3 (ADNI3).[24] Using reports of individuals with ages 70 to 85 years old to match the same age distribution as in UKB, and the fields listed in Table S1, page 2, to categorize AD and controls, 13 AD and 126 controls were selected.

2.2 | Preprocessing and selection of genomic variants

Genome-wide genotyping data from "Affymetrix UK BiLEVE Axiom array" and "Affymetrix UK Biobank Axiom array" available for the 500,000 participants in the UKB cohort was used as the source of genomic data. In the case of ADNI, from the three available GWAS datasets ADNI3 was selected because by using Illumina Infinium Global Screening Arrays was the only one with a specific marker for rs429358 in PLINK files. Bed, bim, and fam files were used to extract individualized genotyping data. Regarding quality assessment, genomic variants with minor allele frequency (MAF) ≤0.01% and Hardy-Weinberg equilibrium (HWE) *P*-value <10E-4 were filtered out. Monomorphic markers (SNVs with the same genotype in all subjects) were also excluded. The list of SNVs to be used as predictors was obtained from the "curated variant disease associations" dataset in DisGeNet,[20] filtering for the AD categories described in Table S1, page 3. A total number of 145 SNVs reported to be related to AD and passing the quality filters mentioned above were selected as AD predictors in UKB to be used in the ML models. The annotated list of AD-related predictors is provided in Table S2 in supporting information and the distribution over chromosomes and genomic regions is shown in Figure S1C, D. Among the AD predictors, 14 SNVs that were prioritized in at least one of the three ML models with UKB and were common in UKB and ADNI3 arrays were selected in ADNI for comparison purposes. After the selection of the variants, numeric matrices were built with rows representing samples and columns representing SNVs. We used the dbSNP ID as unique identifier for SNVs. In the matrix, SNVs were categorized as 1, 2, or 3 corresponding to the three possible genotypes, minor allele of the SNV absent, present in one allele, or present in both alleles, respectively. Missing values were categorized with 0. A quality control based on allele frequency (AF) was applied to discard the presence of major technical artifacts (Table S3 in supporting information).

2.3 | ML models: building and evaluation

Python 3.7.6 with Scikit-learn v0.22.1 module was used to build the ML models on the UKB pre-processed matrix described in the section "Pre-processing and selection of genomic variants." A train/test split was applied to have 80% of samples for training and 20% of samples for testing. Samples were balanced to have the same proportion of LOAD and controls using random undersampling. Nested cross-validation (CV) was applied to discard a significative overfitting in models (Table S4 in supporting information). Hyperparameter selection was performed on the training set through a 10-fold CV. Two metrics such as area under the receiver operating characteristic curve (AUC-ROC) and f-score were considered for determining the best hyperparameter configuration for each model. The median of AUC-ROC (Figure S2A, B, C in supporting information) and f-score (Figure S2D, E, F) was > 0.7 in validation sets of models with AD predictors. The final model with the optimized parameters was trained on the original train set (80%) and tested on the test set (20% of samples). The selected parameters are listed in Table S1, page 4.

2.4 | Statistical test for interactions between pairs of SNVs

For all the possible pairwise combinations of 14 SNVs, R (v4.0.4) was used to build full generalized linear models (glm) considering two SNVs as independent variables with their individual effect and interaction to classify LOAD and controls, and a reduced glm with the same variables but considering only the individual effect of each SNV without the interaction term to classify both classes. The models were built with samples that did not have any missing value in any of the 14 SNVs, consisting in a total number of 616 LOAD and 61,987 controls in UKB, and 12 LOAD and 116 controls in ADNI. The function anova.glm in stats package was used to perform the analysis of deviance between the full glm and reduced glm, comparing the reduction in deviance with a Chi-squared test. In UKB, for each combination of variants the test was applied 1000 times, randomly selecting 500 AD and 500 controls in each iteration to have balanced groups. One hundred iterations subsampling 10 AD and 10 controls were applied in ADNI3. The asymptotically exact harmonic mean *P*-value (HMP) was used to summarize the *P*-values obtained across the iterations and to correct for multiple comparisons.

3 | RESULTS

3.1 | Evaluation metrics across different ML models

We tested the ability of GB, ET, and RF final models with the AD predictors to classify LOAD and controls. The evaluation was made with seven different metrics shown in Table 1. The best scores in all evaluation metrics were obtained using GB with an accuracy, f-score, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of 0.80, and an AUC-ROC of 0.91. RF performed slightly better than ET in accuracy (0.74), f-score (0.74), sensitivity (0.73), specificity (0.75), PPV (0.75), and NPV (0.73), but AUC-ROC was better in ET (0.82) (Figure S3A in supporting information). In agreement with literature,[25] ML models using the *APOE* ε2 and *APOE* ε4 SNVs had a good predictive power, but performances significatively improved when using the set of 145 SNVs (median AUC-ROC of 0.68 against >0.80, Figure S4 in supporting information).

TABLE 1 Summary of the evaluation metrics obtained with GB, ET, and RF models and Alzheimer's Disease predictors

	Accuracy	AUC-ROC	F score	Sensitivity	Specificity	PPV	NPV
GB	0.801	0.912	0.800	0.797	0.804	0.803	0.799
ET	0.707	0.820	0.706	0.703	0.710	0.708	0.705
RF	0.739	0.804	0.735	0.725	0.754	0.746	0.732

Abbreviations: GB, gradient boosted decision trees; ET, extremely randomized trees; RF, random forest; AUC-ROC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value.
Note: Machine learning models with best scores in each evaluation metric are highlighted in red.

TABLE 2 Characteristics of the six SNVs prioritized by the three machine learning methods

SNV	Gene	Region	Chr	hg19 position	AF AD	AF Cntrl	LOG2 FC AF AD/Cntrl	FI RF	FI ET	FI GB	Fisher *P*-value
rs429358	*APOE*	Exonic	19	45411941	0.403	0.156	1.373	0.047	0.041	0.066	3.72E-44
rs769449	*APOE*	Intronic	19	45410002	0.320	0.127	1.329	0.040	0.043	0.119	9.95E-37
rs4420638	*APOC1*	Downstream	19	45422946	0.417	0.189	1.144	0.051	0.050	0.045	4.56E-42
rs405509	*APOE*	Upstream	19	45408836	0.411	0.470	−0.194	0.130	0.152	0.139	1.40E-03
rs1160985	*TOMM40*	Intronic	19	45403412	0.322	0.458	−0.507	0.013	0.026	0.167	5.04E-14
rs7412	*APOE*	Exonic	19	45412079	0.033	0.087	−1.385	0.014	0.015	0.024	5.14E-09

Notes: dbSNP ID together with gene annotations are provided in the columns "SNV," "Gene", "Region", "Chr" and "hg19 position". AF in AD and in controls are used to calculate the log2FC in AD vs. Cntrl (column "LOG FC AF AD/Cntrl"). SNVs are ordered from the highest logFC (top) to the lowest (bottom) and colored in blue and red accordingly. FI obtained in RF, ET, and GB are in columns "FI RF", "FI ET" and "FI GB" respectively. Fisher test *P*-values measuring the significance of AF differences between AD and controls are provided in the "Fisher *P*-value" column.
Abbreviations: AD, Alzheimer's disease; SNV, single nucleotide variants; AF, allele frequencies; *APOE*, apolipoprotein E; FI, feature importance; RF, random forest; ET, extremely randomized trees; GB, gradient boosted decision trees.

3.2 | Prioritization of SNVs using FI

To identify the SNVs that provide the strongest signal for the classification, we ranked the AD predictors based on the impurity-based FI. For each ML method, a FI higher than 0.01 was used to select the SNVs among the 145 in AD predictors considered relevant during the classification (Figure S3B to D). A set of 9, 20, and 15 SNVs were selected in GB, ET, and RF models, respectively (Table S5 in supporting information). We named the SNVs using their dbSNP ID and referring to the presence or absence of the minor allele. Overall, all the prioritized SNVs except one, rs7561528, prioritized by RF in chromosome 2, were located in a region of chromosome 19 comprising *PVRL2*, *TOMM40*, *APOE*, and *APOC1* genes (Figure 1A). The intersection of the prioritized SNVs across the three ML methods is shown in Figure 1B. The six SNVs prioritized by the three ML models were: rs1160985 in *TOMM40*; rs405509, rs7412, rs769449, and rs429358 in *APOE*; and rs4420638 downstream *APOC1* (Table 2).

We used the Fisher-test *P*-value to measure the differences in AF between LOAD and controls and then check if the SNVs with higher FI were also the ones with higher AF differences. Among all AD predictors, the highest AF differences were in a set of SNVs located in chromosome 19 (Figure 1C), in a hot-spot region comprising *PVRL2*, *TOMM40*, *APOE*, and *APOC1* genes, the same region where the prioritized SNVs using FI were located (Figure 1A). Looking at the AF of SNVs prioritized by the three methods in Table 2, SNVs more frequent in LOAD with respect to controls were rs429358 (pval = 3.72E-44 logFC = 1.37), rs769449 (pval = 9.95E-37 logFC = 1.33), and rs4420638 (pval = 4.56E-42 logFC = 1.14). Conversely, the SNVs more frequent in controls were rs7412 (pval = 5.14E-09 logFC = −1.39), rs1160985 (pval = 5.04E-14 logFC = −0.51), and rs405509 (pval = 1.40E-03 logFC = −0.19). rs429358 had the highest AF difference between conditions (Figure 1C), being 2.6 times more frequent in LOAD with respect to controls (logFC = 1.37). Yet, rs429358 was not the SNV with the highest FI in any of the ML methods (Table 2, Table S5). Alternatively, rs405509 reached the highest FI in RF (0.13) and ET (0.15) and the second highest in GB (0.14), but had the lowest AF differences between LOAD and controls compared to the other prioritized SNVs (logFC = −0.19; Figure 1C, Table 2). rs1160985 had the highest FI in GB (0.17) but relatively low differences in AF between LOAD and controls (logFC = −0.507) compared to the other SNVs. We hypothesized the importance of rs405509 and rs1160985 in the ML classification is probably due to the co-occurrence or mutual exclusion with other variants that together form certain genomic profiles, rather than for being more present in one condition with respect to the other.

3.3 | Interactions in the hot-spot region of chromosome 19

To identify possible interactions occurring between the sets of prioritized SNVs, we analyzed the genomic profiles whose samples in UKB were correctly classified as true positives (TP) or true negatives (TN)

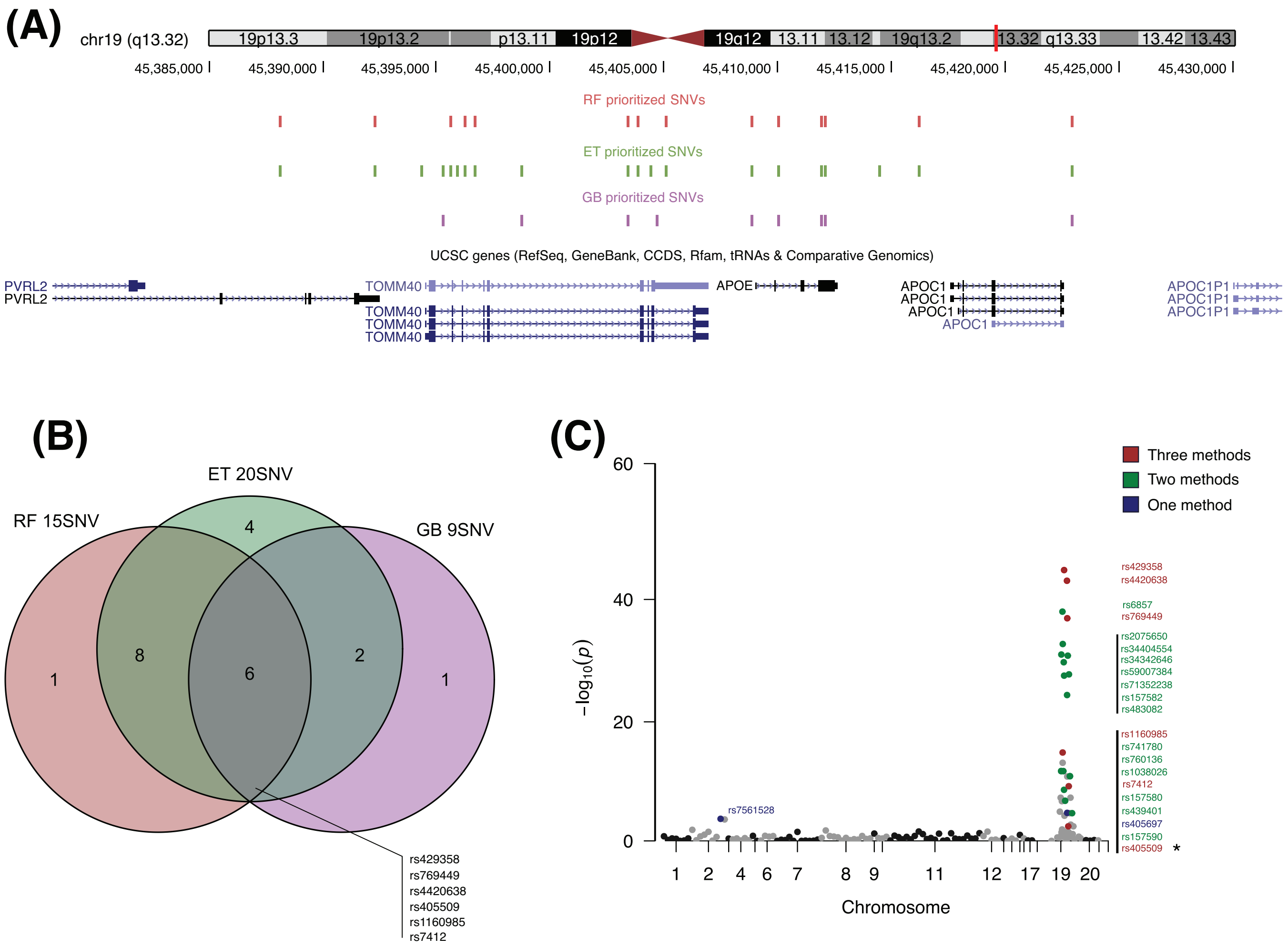

FIGURE 1 A, The genomic location of single nucleotide variants (SNVs) selected using a feature importance (FI) >0.01 in the chromosome 19 hot-spot region. SNVs prioritized by different machine learning (ML) methods are illustrated in different tracks. B, Venn diagram showing the intersection of the prioritized SNVs by gradient boosted decision trees (GB), extremely randomized trees (ET), and random forest (RF). The name of the SNVs in the intersection with the three methods is provided. C, For the 145 SNVs in Alzheimer's disease (AD) predictors, distribution of the Fisher-test *P*-values obtained measuring differences in allele frequency (AF) between late onset Alzheimer's disease (LOAD) and controls over the chromosomes. The name of the SNVs prioritized by any of the three ML methods is provided and a color is assigned depending on the number of times a SNV was selected by any one of the methods. The six SNVs prioritized by GB, ET, and RF are colored in red

in GB, ET, and RF (Figure 2, Figure S5 in supporting information, and Figure S6 in supporting information, respectively). Most of the patterns described hereafter are observed in genomic profiles captured by the three ML methods. However, for simplicity we discuss the genomic profiles defined by GB only (Figure 2). This decision is supported by the fact that GB was the model with the best performance in the classification.

Most TP profiles were characterized as having the three SNVs, rs429358, rs4420638, and rs769449 either in one or two alleles (Figure 2A, C1). These genomic profiles were present in the 27.83% of LOAD and 1.21% of controls in the full UKB dataset. The genomic profiles with rs405509 in both alleles co-occurring with rs1160985 were present in 12.62% of LOAD and were not present in controls (Figure 2A, C3). Interestingly, rs405509 in both alleles and rs1160985 were mutually exclusive in TN (Figure 2B, C1 and C2). Moreover, the predisposition to AD caused by the presence of rs429358, rs4420638, and rs769449 in one allele was neutralized with the presence of rs405509 in both alleles in a group of TN (Figure 2B, C1). Only GB captured profiles of TP characterized with rs405697 and rs157580 co-occurring either in one or two alleles without the presence of the other seven SNVs in 3.24% of LOAD (Figure 2A, C2).

Altogether, the LOAD genomic profiles captured by ML models suggested that an interaction may exist between the SNVs in the chromosome19 hot-spot region. To test the significance of the interactions with an alternative method, we built generalized linear models (glm) to classify LOAD and controls with and without the interaction term using pairwise combinations of SNVs and measured the changes in the deviance between both models. We evaluated the interactions also in ADNI to compare the results to an external dataset. We applied the test on the 14 SNVs prioritized by any of the three ML methods with UKB that were also present in ADNI3 arrays (Figure 3). In accordance with the patterns in genomic profiles, rs405509 and rs1160985 were enriched in statistically significant pairwise interactions with the

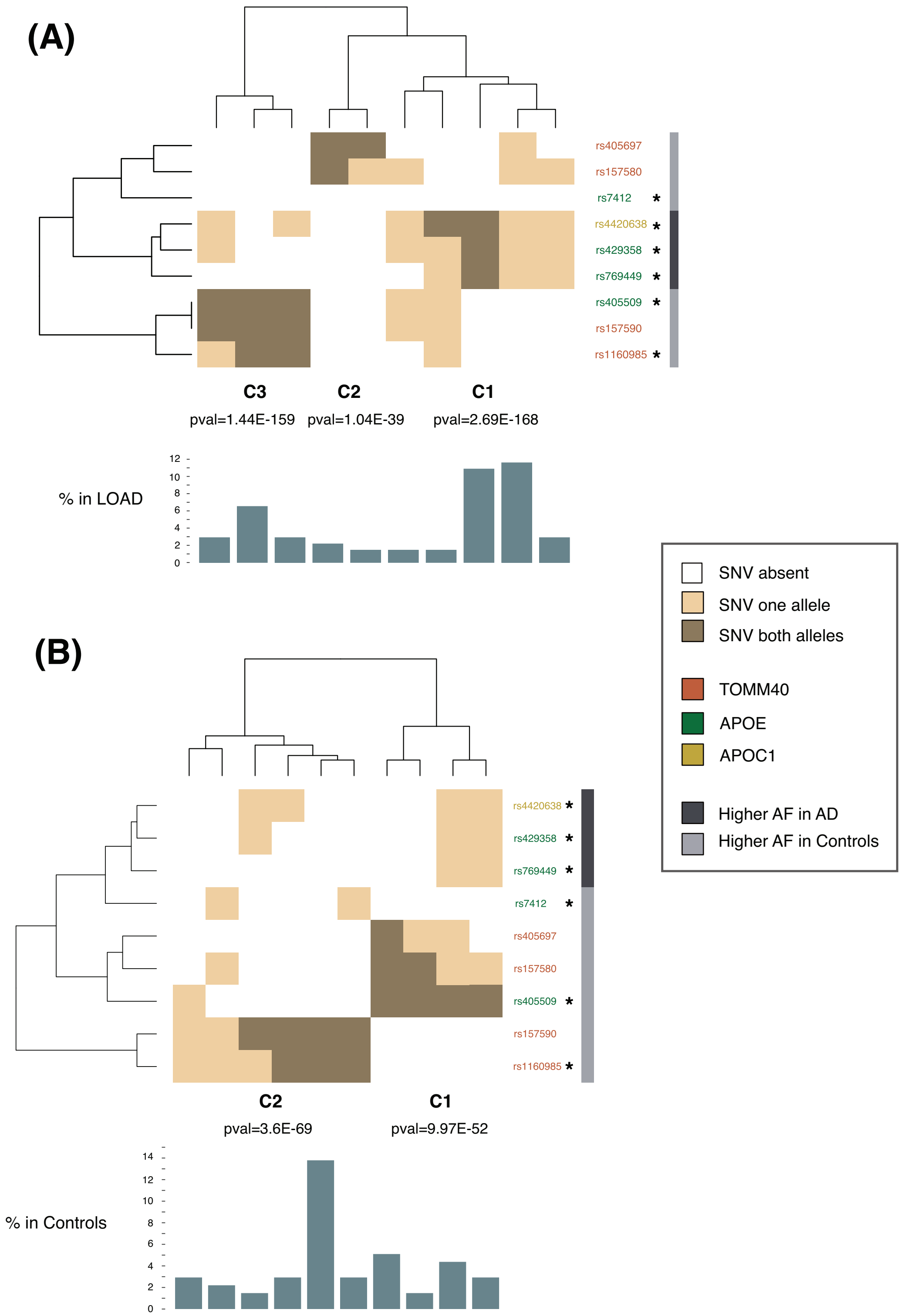

FIGURE 2 Genomic profiles of correctly classified samples in gradient boosted decision trees (GB) defined with the nine prioritized single nucleotide variants (SNVs). Genomic profiles with only one sample or having missing values were excluded. In (A) genomic profiles of true positives (TP) represent all samples that were correctly classified as late onset Alzheimer's disease (LOAD). In (B) genomic profiles of true negatives (TN) represent all samples that were correctly classified as controls. Dendrograms on the top and the left were made with Ward-D2 method and Euclidean distances. Clusters of genomic profiles are indicated with numbers in the x-axis. Fisher-test *P*-values are provided measuring the statistical significance of different representation of Alzheimer's disease (AD) and controls in clusters of genomic profiles. The % of samples having each genomic profile in LOAD and controls is indicated in the bar-plots below the heatmaps. SNVs are colored with their corresponding gene loci and information of the higher allele frequency (AF) in LOAD or controls is provided in the right-side bar. An asterisk points to the six SNVs commonly prioritized by GB, extremely randomized trees (ET), and random forest (RF).

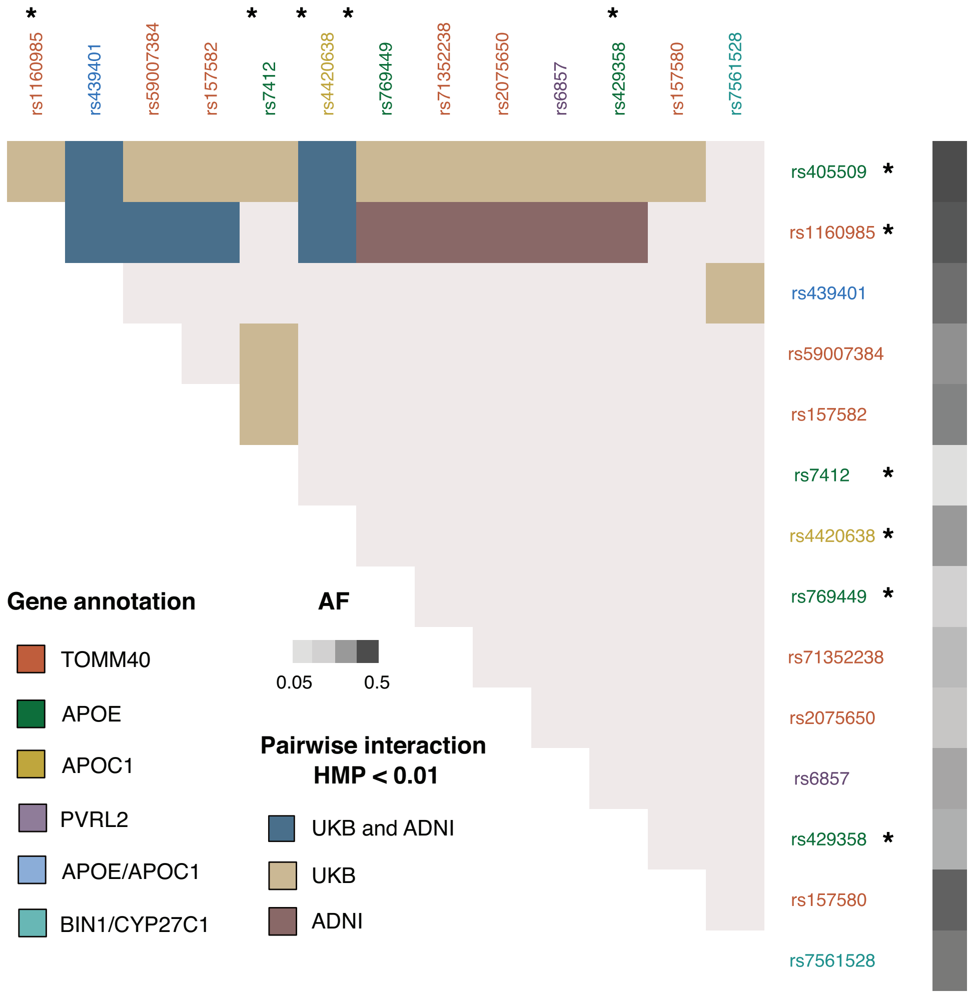

FIGURE 3 Representation of the pairwise test of interactions between the 14 single nucleotide variants (SNVs) prioritized by any of the three machine learning (ML) methods and commonly present in UK Biobank (UKB) and Alzheimer's Disease Neuroimaging Initiative 3 (ADNI3) arrays. Further details on the approach used to test the statistical significance of the pairwise interactions are provided in the Methods section "Statistical test for interactions between pairs of single nucleotide variants (SNVs)". Details on the asymptotically exact harmonic mean *P*-values (HMP) for each pairwise interaction are provided in Table S6. A cut-off HMP <0.01 was used to consider an interaction statistically significant. An asterisk points to the six SNVs commonly prioritized by gradient boosted decision trees (GB), extremely randomized trees (ET), and random forest (RF). Statistically significant interactions are enriched with rs1160985 and rs405509 in both datasets. These two SNVs: (1) had high feature importance (FI) scores in the ML models, (2) had low allele frequency (AF) differences between Alzheimer's disease (AD) and controls, (3) were involved in interaction patterns of the genomic profiles obtained with ML approaches. In UKB, 16 of the 19 statistically significant pairwise interactions involved rs1160985 or rs405509 (Fisher test *P*-value 5.28E-09). In ADNI all the statistically significant pairwise interactions involved one of the two SNVs (Fisher test *P*-value 9.42E-08). SNVs are ordered from the top to the bottom and from the left to the right by number of statistically significant interactions (decreasing). The gray gradient corresponding to the AF shows weak correlation between number of statistically significant pairwise interactions of SNVs and AF (spearman correlation 0.41 and 0.32 in UKB and ADNI, respectively)

other prioritized SNVs in ADNI and UKB. In UKB 16 of the 19 statistically significant pairwise interactions involved rs1160985 or rs405509 (Fisher *P*-value 5.28E-09). In ADNI all the statistically significant pairwise interactions involved one of the two SNVs (Fisher *P*-value 9.42E-08). There were 12 statistically significant interactions with rs405509 and 10 with rs1160985, from which six were common in UKB and ADNI (Figure 3 and Table S6 in supporting information). The enrichment of interactions in these SNVs cannot be attributed to the high AF. For example, rs157580 also had AF >0.39, but was involved in only one significant interaction, while rs7412 had AF <0.1 and had three significant interactions.

3.4 | Expression and splicing quantitative trait loci in the prioritized SNVs

We examined the effect of the six SNVs commonly prioritized by the three ML methods on gene expression (expression quantitative trait loci, eQTL) and splicing (splicing quantitative trait loci, sQTL) on different tissues in GTEx.[26] Regarding gene expression, rs1160985, rs405509, rs7412 and rs4420638 were eQTLs of *APOE* in skin tissue and in the case of rs1160985 heart tissue as well (Table S1, page 6). rs429358, rs769449, and rs4420638 were eQTLs of the upstream gene *APOC1* in esophagus, adrenal gland, and skin. Even if there were no eQTLs captured in brain tissue for any of the six SNVs, these data evidence the presence of a transcriptional regulatory hub in the hotspot region of chromosome 19 that may be altered by the presence of alternative alleles in the prioritized SNVs. On the other side, all the SNVs except for rs7412 were sQTLs to *TOMM40* in the brain (Table S1, page 7). In addition, rs405509 and rs429358 were sQTL to *APOE* in lung and brain, respectively.

4 | DISCUSSION

Using tree-based ML methods and the set of 145 SNVs related to AD reported in databases, we were able to classify LOAD and controls, reaching an accuracy of 0.80 and an AUC-ROC of 0.91 in GB. We prioritized a set of 9, 20, and 15 SNVs in GB, ET, and RF, respectively, from which six SNVs were commonly prioritized across the three methods. The six SNVs were located in a chromosome19 hot-spot region comprising *TOMM40*, *APOE*, and *APOC1* genes. rs429358 is the most well-characterized LOAD genetic determinant,[27,9] rs7412 is known to be protective against AD,[28] and the two SNVs define the distinct *apoE* isoforms.[25] rs4420638 is in strong LD with rs429358[8] and for this reason, its link with AD is attributed to rs429358. rs769449 has been associated with low-density lipoprotein cholesterol plasma levels,[29] with lower longevity,[30] and with cognitive decline.[31] rs1160985 has been related to increased risk of LOAD in a Chinese population,[32,33] but to be protective against AD in other ethnic cohorts.[34,35] Located in the *APOE* promoter region, the rs405509 minor allele in both copies has been described to alter *APOE* gene expression[36,37] and to act as effect modifier to rs429358 in previous AD studies.[38,39,40] Intriguingly, the SNVs reaching the highest FI, rs405509 and rs1160985, had relatively low AF differences between LOAD and controls compared to the other prioritized SNVs. Also, the two most well-characterized LOAD genetic determinants in the literature to date and the ones with higher AF differences between both conditions, rs429358[27] and rs7412,[28] were not the ones with the highest FI scores. These results suggest that tree-based ML methods are capable of prioritizing variants not only based on the individual enrichment of each SNV in the different classes, but also considering interactions between groups of SNVs.

Looking at the correctly classified genomic profiles, most of the TP were characterized to have rs429358, rs4420638, and rs769449 co-occurring in either one or two alleles, without the presence of rs405509 in two alleles and the absence of rs7412. Contrarily, profiles with rs429358, rs4420638, and rs769449 in one allele co-occurring with rs405509 in two alleles were present in TN. In this sense, rs405509 seems to act as an effect modifier over the three predisposing variants. In addition, rs1160985 and rs405509 in both alleles were either predisposing to AD when co-occurring in TP, or protecting against AD when mutually exclusive in TN. Last, rs157580 and rs405697 were present in a small number of LOAD cases, comprising a third group of TP in GB. These two SNVs in *TOMM40* were reported in other works to be related with lower longevity in the Chinese population[41] and related to AD independently of variants in the *APOE* locus in the Japanese population.[35] Testing the statistical significance of the pairwise interactions in the set of 14 SNVs prioritized by any of the three ML methods, results corroborated what we observed in the genomic profiles. There was an enrichment of rs405509 and rs1160985 in the statistically significant interactions in UKB and ADNI datasets.

Using GTEx data we show that the six prioritized SNVs are eQTLs of *APOE* or *APOC1*, but not in brain tissues. Conversely, rs1160985, rs405509, rs769449, rs4420638 are sQTLs of *TOMM40*, and rs429358 is sQTL of *TOMM40* and *APOE* in brain tissues. In this respect, some studies previously suggested the existence of a complex transcriptional regulatory hub in the region where the prioritized SNVs are located.[10,42,43] The three predisposing SNVs prioritized by the three ML methods (rs429358, rs4420638, and rs769449) are in high LD. Among the three, only rs4420638 was an eQTL to *APOE*. Seemingly, only rs429358 was sQTL to *APOE*, possibly indicating different biological mechanisms despite LD. However, with the data we had we could not demonstrate that the three SNVs were independently associated with AD. Other studies indicated that ML methods are robust in terms of performance when dealing with SNVs in LD.[44,45,46]

As in other works,[16] we found that tree-based ML methods can add an important layer of information to the disease-related variants obtained with other population genomic approaches such as GWAS. We proved that ML methods are efficient at classifying the major genetic profiles defined by a set of interactions between SNVs. The validation of these genomic profiles could improve the clinical characterization of patients in the future. Nevertheless, the possibility of using individualized genomic information to stratify the population with the risk of developing a certain disease, especially if a cure is not yet available, is always controversial. With the balance of benefits and costs in mind, genetic tests could further the health-care system implementing preventive measures in a healthy population with the risk of developing AD. Yet, an adequate regulation should be applied, considering topics such as personal data protection, privacy, and informed consent.[47]

ACKNOWLEDGMENTS

We thank all members of Tartaglia, Paliouras and Garrard's labs. The research leading to these results was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 727658 (project iASiS), European Research Council ASTRA 855923, and the European Genome-phenome Archive (EGA). Claudia Giambartolomei has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 754490–MINDED project. This research has been conducted using the UK Biobank Resource under Application Number 35916.

CONFLICTS OF INTEREST

Support received for the submitted work: Jordi Rambla De Argila, George Paliouras, Peter Garrard, and Gian Gaetano Tartaglia received funds through the H2020 project IASIS (grant number 727658 to CRG, NCSR, and St George's University of London); Claudia Giambartolomei is recipient of a H2020 Skłodowska-Curie fellowship (grant number 754490 to IIT). The other authors have nothing to disclose.

Support received outside the submitted work: Gian Gaetano Tartaglia received funds through the ERC Synergy project ASTRA (grant number 855923 to IIT and Sapienza). Peter Garrard received funds from Medical Research Council UK and Leverhulme Trust (to St George's University of London). The other authors have nothing to disclose. Other relationships/activities/interests: Gian Gaetano Tartaglia is in the process of filing a patent on TDP43 aptamers and received an honorarium for editorial work at Frontiers in Molecular Biosciences. George Paliouras filed the following patents: EPO Patent Application 20 386 008.5 - 1126: System and method for identifying drug-drug interactions USPTO Patent Application US-2021-0118465-A1: Multilingual System for Early Detection of Alzheimer's Disease EPO Patent Application EP-3- 809-411-A1: Multi-lingual System for Early Detection of Alzheimer's Disease. Other relationships with entities: George Paliouras participated on a Data Safety Monitoring/Advisory Board in the last 36 months for the EIT High Value Care Forum (no payment); George Paliouras held a leadership/fiduciary role in the Duchenne Data Foundation (no payment). The other authors have nothing to disclose.

ORCID

Magdalena Arnal Segura https://orcid.org/0000-0001-8647-9232

REFERENCES

1. Ballard C, Gauthier S, Corbett A, et al. Alzheimer's disease. *Lancet*. 2011;377:1019-1031.
2. Cacace R, Sleegers K, Van Broeckhoven C. Molecular genetics of early-onset Alzheimer's disease revisited. *Alzheimer's Dement*. 2016;12:733-748.
3. Rabinovici GD. Late-onset Alzheimer disease. *Continuum*. 2019;25:14-33.
4. Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet*. 2019;51:414-430.
5. Tam V, Patel N, Turcotte M, et al. Benefits and limitations of genome-wide association studies. *Nat Rev Genet*. 2019;20:467-484.
6. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*. 2018;19:581-590.
7. Gatz M, Reynolds CA, Fratiglioni L, et al. Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry*. 2006;63:168-174.
8. Coon KD, Myers AJ, Craig DW, et al. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J Clin Psychiatry*. 2007;68:613-618.
9. Huang Y-WA, Zhou B, Nabet AM, Wernig M, Südhof TC. Differential signaling mediated by ApoE2, ApoE3, and ApoE4 in human neurons parallels Alzheimer's disease risk. *J Neurosci*. 2019;39:7408-7427.
10. Zhou X, Chen Y, Mok KY, et al. Non-coding variability at the APOE locus contributes to the Alzheimer's risk. *Nat Commun*. 2019;10:1-16.
11. Escott-Price V, Sims R, Banniste C, et al. Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain*. 2015;138:3673-3684.
12. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine learning SNP based prediction for precision medicine. *Front Genet*. 2019;10;267.
13. Bracher-Smith M, Crawford K, Escott-Price V. Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Mol Psychiatry*. 2021;26:70-79. https://doi.org/10.1038/s41380-020-0825-2
14. Romero-Rosales BL, Tamez-Pena JG, Nicolini H, Moreno-Treviño MG, Trevino V. Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modeling. *PLoS One*. 2020;15:e0232103.
15. De Velasco Oriol J, Vallejo EE, Estrada K, Taméz Peña JG & Disease Neuroimaging Initiative, T. A. s. Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data. *BMC Bioinformatics*. 2019;20:709.
16. Nicholls HL, John CR, Watson DS, et al. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. *Front Genet*. 2020;11:350.
17. Vasilopoulou C, Morris AP, Giannakopoulos G, Duguez S, Duddy W. What can machine learning approaches in genomics tell us about the molecular basis of amyotrophic lateral sclerosis? *J Pers Med*. 2020;10:1-28.
18. Wang Y, Goh W, Wong L, Montana G & Initiative, the A. D. N.. Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes. *BMC Bioinformatics*. 2013;14:S6.
19. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12:e1001779.
20. Piñero J, Bravo A, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*. 2017;45:D833–D839.
21. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42:D980-D985.
22. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47:D1005-D1012.
23. Li MJ, Liu Z, Wang P, et al. GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res*. 2016;44:D869-D876.
24. Saykin AJ, Shen L, Yao X, et al. Genetic studies of quantitative MCI and AD phenotypes in ADNI: progress, opportunities, and plans. *Alzheimers Dement*. 2015;11(7):792-814.
25. Raber J, Huang Y, Ashford JW. ApoE genotype accounts for the vast majority of AD risk and AD pathology. *Neurobiol Aging*. 2004;25:641-650.
26. LJ C, Ardlie K, Barcus M, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx Project. *Biopreserv Biobank*. 2015;13:311-317.
27. Corder EH, Saunders AM, Strittmatter WJ, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science (80-.)*. 1993;261:921-923.
28. Corder EH, Saunders AM, Risch NJ, et al. Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nat Genet*. 1994;7:180-184.
29. Chasman DI, Paré G, Zee RYL, et al. Genetic loci associated with plasma concentration of low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, apolipoprotein A1, and Apolipoprotein B among 6382 white women in genome-wide analysis with replication. *Circ Cardiovasc Genet*. 2008;1:21-30.
30. Soerensen M, Dato S, Tan Q, et al. Evidence from case–control and longitudinal studies supports associations of genetic variation in APOE, CETP, and IL6 with human longevity. *Age (Omaha)*. 2013;35:487-500.
31. Zhang C, Pierce BL. Genetic susceptibility to accelerated cognitive decline in the US Health and Retirement Study. *Neurobiol Aging*. 2014;35:1512.e11-1512.e18.
32. Jiao B, Liu X, Zhou L, et al. Polygenic analysis of late-onset Alzheimer's disease from mainland China. *PLoS One*. 2015;10:e0144898.
33. Ma XY, Yu J-T, Wang W, et al. Association of TOMM40 polymorphisms with late-onset Alzheimer's disease in a northern han chinese population. *NeuroMol Med*. 15:279-287.

34. Roses AD, Lutz MW, Amrine-Madsen H, et al. A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. *Pharmacogenomics J*. 2010;10:375-384.
35. Takei N, Miyashita A, Tsukie T, et al. Genetic association study on in and around the APOE in late-onset Alzheimer disease in Japanese. *Genomics*. 2009;93:441-448.
36. Laws SM, Hone E, Gandy S, Martins RN. Expanding the association between the APOE gene and the risk of Alzheimer's disease: possible roles for APOE promoter polymorphisms and alterations in APOE transcription. *J Neurochem*. 2003;84:1215-1236.
37. Lescai F, Chiamenti AM, Codemo A, et al. An APOE haplotype associated with decreased 4 expression increases the risk of late onset Alzheimer's disease. *J Alzheimer's Dis*. 2011;24:235-245.
38. Choi K, Lee JJ, Gunasekaran TI, et al. APOE promoter polymorphism-219T/G is an effect modifier of the influence of APOE ε4 on Alzheimer's disease risk in a multiracial sample. *J Clin Med*. 2019;8:1236.
39. Ma C, Zhang Y, Li X, et al. Is there a significant interaction effect between apolipoprotein E rs405509 T/T and ε4 genotypes on cognitive impairment and gray matter volume? *Eur J Neurol*. 2016;23:1415-1425.
40. Bizzarro A, Seripa D, Acciarri A, et al. The complex interaction between APOE promoter and AD: an Italian Case-Control Study. *Eur J Hum Genet*. 2009;17:938-945.
41. Lin R, Zhang Y, Yan D, et al. Association of common variants in TOMM40/APOE/APOC1 region with human longevity in a Chinese population. *J Hum Genet*. 2016;61:323-328.
42. Walker RM, Vaher K, Bermingham ML, et al. Identification of epigenome-wide DNA methylation differences between carriers of APOE ε4 and APOE ε2 alleles. *Genome Med*. 2021;13:1.
43. Bekris LM, Lutz F, Yu CE. Functional analysis of APOE locus genetic variation implicates regional enhancers in the regulation of both TOMM40 and APOE. *J Hum Genet*. 2012;57:18-25.
44. Meng YA, Yu Y, Cupples LA, Farrer LA, Lunetta KL. Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinformatics*. 2009;10:78.
45. Monk B, Rajkovic A, Petrus S, et al. A machine learning method to identify genetic variants potentially associated with Alzheimer's disease. *Front Genet*. 2021;12:642.
46. Motsinger AA, Reif DM, Fanelli TJ, Davis AC, Ritchie MD. Linkage disequilibrium in genetic association studies improves the performance of grammatical evolution neural networks. *Proc IEEE Symp Comput Intell Bioinform Comput Biol*. 2007;2007:1.
47. Ienca M, Vayena E, Blasimme A. Big data and dementia: charting the route ahead for research, ethics, and policy. *Front Med*. 2018;5:13.
48. machalen/ML_LOAD_SNVs. https://github.com/machalen/ML_LOAD_SNVs
49. Genomic search. https://biobank.ctsu.ox.ac.uk/crystal/gsearch.cgi

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Arnal Segura M, Bini G, Fernandez Orth D, et al. Machine learning methods applied to genotyping data capture interactions between single nucleotide variants in late onset Alzheimer's disease. *Alzheimer's Dement*. 2022;14:e12300. https://doi.org/10.1002/dad2.12300