www.nature.com/scientificreports

SCIENTIFIC REPORTS

natureresearch

OPEN

Prediction of Alzheimer's disease using blood gene expression data

Taesic Lee[1] & Hyunju Lee[1,2,3*]

Identification of AD (Alzheimer's disease)-related genes obtained from blood samples is crucial for early AD diagnosis. We used three public datasets, ADNI, AddNeuroMed1 (ANM1), and ANM2, for this study. Five feature selection methods and five classifiers were used to curate AD-related genes and discriminate AD patients, respectively. In the internal validation (five-fold cross-validation within each dataset), the best average values of the area under the curve (AUC) were 0.657, 0.874, and 0.804 for ADNI, ANMI, and ANM2, respectively. In the external validation (training and test sets from different datasets), the best AUCs were 0.697 (training: ADNI to testing: ANM1), 0.764 (ADNI to ANM2), 0.619 (ANM1 to ADNI), 0.79 (ANM1 to ANM2), 0.655 (ANM2 to ADNI), and 0.859 (ANM2 to ANM1), respectively. These results suggest that although the classification performance of ADNI is relatively lower than that of ANM1 and ANM2, classifiers trained using blood gene expression can be used to classify AD for other data sets. In addition, pathway analysis showed that AD-related genes were enriched with inflammation, mitochondria, and Wnt signaling pathways. Our study suggests that blood gene expression data are useful in predicting the AD classification.

Alzheimer's disease (AD), the most common form of dementia, is estimated to affect in 13.8 million individuals in the United States (US), with 7.0 million being aged 85 years or older by 2050[1]. Based on the National Institute of Neurological, Communicative Disorders, and Stroke and Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria in 1985, probable or possible AD was diagnosed based on subjective symptoms and questionnaires[2]. Recently, the transition from symptom-based to pathophysiology-based AD diagnosis showed that AD diagnosis is mainly based on structural brain changes (MRI), molecular neuroimaging changes (positron emission tomography imaging), and alterations in cerebral spinal fluid biomarkers[3]. Although the elucidation of the biological basis of AD has resulted in many advancements[3], early diagnostic detection of AD remains challenging.

Recent advances in biotechnology have led to full-scale analyses of the genome, transcriptome, and epigenome rather than focusing on a few biomarkers. A large-scale genome-wide association study (GWAS) of 2,032 individuals with AD and 5,328 controls was presented in 2009 and it identified variants at CLU and CR1, which were associated with AD[4]. Additionally, a meta-analysis of four previously reported GWAS datasets (17,008 AD cases, 37,154 controls) yielded 11 new loci of susceptibility to AD[5]. Recently, Xu *et al.* constructed an AlzData database integrating data from GWAS, eQTL, interactome, and laboratory experiments[6], which provides all human genes with scores for association with AD, called the Convergent Functional Genomics (CFG) score[7,8].

In recent years, two large multi-center studies were conducted to identify biomarkers for early AD diagnosis and MCI progression to AD: the Europe-based ANM and US-based Alzheimer's Disease Neuroimaging Initiative (ADNI)[9,10]. Furthermore, a large amount of publicly available gene expression data on AD have been provided in the NCBI GEO[11]. As a result, various studies, especially gene expression-based studies, have been published to uncover the informative genes associated with AD. The BrainNet study analyzed 113 samples of well-characterized postmortem brain tissues, yielding 21 genes dysregulated in AD cases[12]. A study by Liang *et al.*[13] consisting of 87 brain tissues samples revealed that in brain tissues of AD cases, the genes encoding subunits of the mitochondrial components showed significantly lower expression. By analyzing RNA expression from brain tissues of AD patients, Xu *et al.*[6] demonstrated that an early alteration of YAP1 could promote AD.

Although several studies involving gene expression data have uncovered valuable patterns, most gene expression data were obtained from biopsy or autopsy-based samples, which are difficult to extrapolate to clinical settings. Only a few studies used blood-based expression data for uncovering key genes related to AD or predicting

[1]Department of Biomedical Science and Engineering, Gwangju Institute of Science and Technology, Gwangju, South Korea. [2]Artificial Intelligence Graduate School, Gwangju Institute of Science and Technology, Gwangju, South Korea. [3]School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, South Korea. *email: hyunjulee@gist.ac.kr

Study	Data source/# of AD and CN (Training and test datasets)	Feature selection methods (Data used for feature selection)	Classifying method	Number of selected features	Performance
Booij *et al.*[15]	Not publicly available (Norway)/126 AD and 126 CN (Randomly dividing all data into training and test datasets by 3:1 ratio)	Jack-knife (training data)	PLSR	1239 genes	ACC: 0.87 AUC: 0.94
Lunnon *et al.*[16]	ANM/104 AD and 104 CN (Randomly dividing AD and CN data into training and test datasets by 3:1 ratio)	t-test RF with Meng score and backward elimination (training data)	RF	50 probes	ACC: 0.75
Sood *et al.*[17]	ANM1 and ANM2/49 AD and 64 CN, 40 AD and 71 CN (LOOCV)	Bayesian statistic (ULSAM Ageing data GEO:GSE60862)	kNN	150 probes	AUC: 0.73 (ANM1) AUC: 0.66 (ANM2)
Voyle *et al.*[18]	ANM1 and ANM+DCR/100 AD and 107 CN, 118 AD and 118 CN (ANM1 for training, ANM2 + DCR for test)	REF and pickSizeTolerance (Training data)	RF	13 probes (12 genes)	ACC: 0.657 AUC: 0.724
Li *et al.*[19]	ANM1 and ANM2/145 AD and 104 CN, 140 AD and 135 CN (ANM1 for training, ANM2 for test and vice versa)	Ref-REO (Training data)	Not described	1,145 gene pairs (ANM1: training data) 1,249 gene pairs (ANM2: training data)	AUC: 0.733 (ANM2: test set) AUC: 0.775 (ANM1: test set)
Li *et al.*[20]	ANM1 and ANM2/143 AD and 104 CN, 102 AD and 78 CN (ANM1 for training, ANM2 for test and vice versa)	LASSO regression (ANM1 and ANM2)	Majority voting of SVM, RR and RF	6 genes (Full6set)	AUC: 0.866 (ANM2: test set) AUC: 0.864 (ANM1: test set)

Table 1. Summary of six studies for predicting AD using blood gene expression data. AD: Alzheimer's Disease; CN: healthy control; PLSR: partial least square regression; ACC: accuracy; AUC: area under the curve; ANM: AddNueroMed; RF: Random Forest; kNN: k-nearest neighbors; RFE: recursive feature elimination; pickSizeTolerance: a function in caret package[29]; ULSAM: the Uppsala Longitudinal Study of Adult Men; LOOCV: leave-one-out cross-validation; LASSO: least absolute shrinkage and selection operator; SVM: support vector machine; RF: random forest; RR: logistic ridge regression.

early AD[14]. Cooper *et al.*[14] published a study consisting of 186 AD cases, 118 MCI cases, and 204 controls from three independent datasets, overall suggesting that progranulin expression levels in the blood are increased in AD and MCI.

In Table 1, we list previously published blood expression-based studies for the identification of patients with AD, especially focusing on machine learning (ML)-based studies[15–20]. Detailed information about related works is presented in Supplementary File. The difference between the studies outlined in Table 1 and ours are as follows. First, other studies selected AD-related genes only using statistical or ML methods. We extracted AD-related genes not only via statistical methods, but also protein-protein interaction databases (DBs), transcription factor (TF) DBs, and the CFG method that integrated results from SNP, transcripts, AD-animal model, and text-mining. Besides, we validated selected genes by measuring predictive performances among different datasets. Although studies discriminating AD by blood-based transcriptomic data have been performed, it is unclear whether the classifiers trained using one AD gene expression data set can be applied to other AD data sets. Thus, here, we systematically evaluated five feature selection methods and classifiers for distinguishing individuals with AD from healthy controls (CNs) using three independent blood gene expression datasets. Lastly, we analyzed the biological functions of AD-related genes from the blood via pathway analysis, and compared the result from blood bio-signature with those from brain bio-signature.

Methods

The ADNI consisted of participants recruited at 57 sites in the US and Canada, funded as a private-public partnership[9]. The ANM consortium is a large cross-European AD biomarker study and a follow-on DCR cohort in London[10]. In both ADNI and ANM, AD was diagnosed using the NINCDS-ADRDA criteria for possible or probable AD[2].

We employed three large-scale blood gene expression datasets: ANM1 (GEO:GSE63060), ANM2 (GEO:GSE63061), and ADNI (adni.loni.usc.edu, last downloaded 2018/8/31). The overall framework of our study is illustrated in Fig. 1. The performance was evaluated via internal validation (five-fold CV within each dataset) and external validation (training and test sets from different datasets). In the internal validation. Detailed information about performance assessment is described in the Supplementary File.

The extraction of differentially expressed genes (DEGs), logistic regression (LR), L1-LR, support vector machine (SVM), and random forest (RF) were implemented using R (version 3.4.0) and Bioconductor (release 3.8)[21,22]. Variational autoencoder (VAE) and deep neural network (DNN) were implemented with C++ based TensorFlow with a Python interface[23].

Preprocessing of data.

Gene expression data were produced using the Affymetrix Human Genome U 219 array for the first case-control study (ADNI), Illumina Human HT-12 v.3 Expression BeadChips for the second

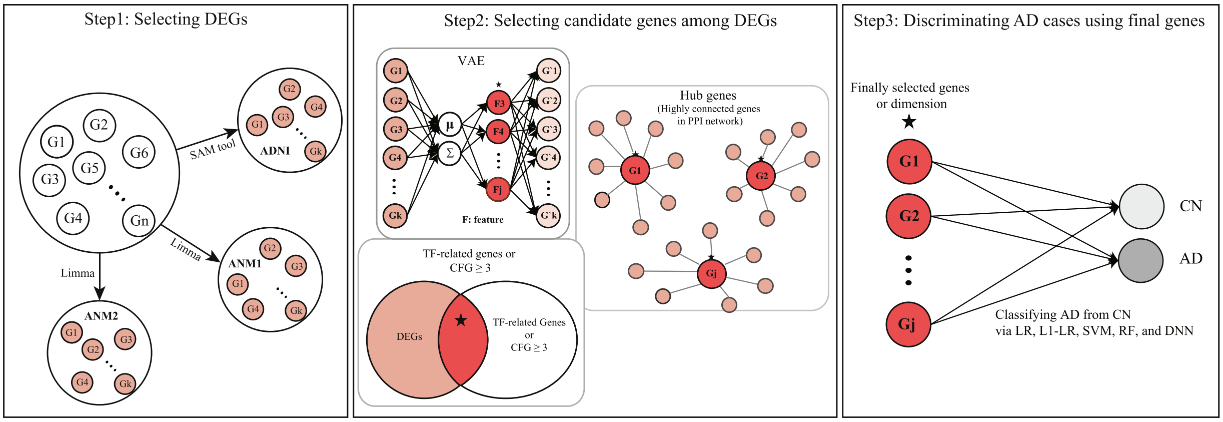

Step1: Selecting DEGs
G1
G2
G6
G3
G5
G4
Gn
SAM tool
ADNI
Limma
Limma
ANM1
ANM2
Step2: Selecting candidate genes among DEGs
VAE
μ
Σ
F3
F4
Fj
F: feature
G'1
G'2
G'3
G'4
G'k
Hub genes
(Highly connected genes in PPI network)
TF-related genes or CFG ≥ 3
DEGs
TF-related Genes or CFG ≥ 3
Step3: Discriminating AD cases using final genes
Finally selected genes or dimension
G1
G2
Gj
CN
AD
Classifying AD from CN via LR, L1-LR, SVM, RF, and DNN

Figure 1. A framework of the study. The informative genes were selected using a training set by two processes: (Step 1) Extracting DEGs; (Step 2) Selecting informative genes using feature selection methods, including VAE, TF-related genes, hub genes, and the CFG scoring; (Step 3) Learning a prediction model using a training set, and predicting a test set by employing five classification methods, including logistic regression (LR), L1-regularized LR (L1-LR), SVM, RF, and DNN. Stars denote the best performance among five classifying methods. DEG, differentially expressed gene; SAM, significance analysis of microarray; ADNI, Alzheimer's Disease Neuroimaging Initiative; ANM, AddNeuroMed; VAE, variational autoencoder; TF, transcription factor; PPI, protein-protein interaction; CFG, Convergent Functional Genomics; LR, logistic regression; SVM, support vector machine; RF, random forest; DNN, deep neural network; CN, healthy control; AD, Alzheimer's disease.

case-control study (ANM1), and Illumina Human HT-12 v4 Expression BeadChips for the third case-control study (ANM2).

We performed three steps of data processing. The first step involved the selection of samples and probes to be analyzed. For ADNI, we selected high-quality RNA samples with RIN $\geq$ 6.9, as the previous study performed[24], and for ANM1 and ANM2, we did not exclude any samples. The ADNI, ANM1, and ANM2 datasets consisted of 49,386, 30,063, and 29,485 probes, respectively. The median RNA expression value of ADNI, ANM1, and ANM2 were 3.897, 7.584, and 6.154, respectively, indicating that ADNI could be influenced by background noise due to relatively low gene expression intensities. To reduce the background noise of ANDI, we excluded probes with an intensity value $\leq$ the median of all gene expression values in 100 or more samples, as performed in the previous study[25]. If there were multiple probes annotated in one gene, then the median value of those was selected, yielding 11,276, 21,698, and 22,338 unique probes in ADNI, ANM1, and ANM2, respectively. We selected only probes that were present in all three datasets, and 8,835 final probes were left for analysis.

The second step was normalization within each dataset and renormalization between datasets. The probe set level intensities of all three datasets were normalized by the Robust Multi-Array Analysis (RMA) method[26]. Although each dataset was normalized, a variance or batch effect among different datasets remained. Therefore, we renormalized all three datasets to reduce the batch effect among different datasets using ComBat from the sva package in R[27].

The third step was selecting DEGs of patients with AD. We extracted DEGs between the control and AD in the ADNI by the significance analysis of microarrays (SAM), which is a method for identifying genes on a microarray with statistically significant changes in expression[28]. It was developed in the context of an actual biological experiment. DEGs of AD in ANM1 and ANM2 were curated via "lmFit" and "eBayes" functions in the limma package, which is based on a linear regression method[29]. We set the cutoffs of FDR of the SAM and limma to 0.05 and 0.01, respectively.

Feature selection. We used VAE to extract a representation from a set of input features thereby reducing the dimensions of the data[30]. An autoencoder is a type of neural network used to learn efficient and representative information in an unsupervised manner. Specifically, VAE not only adopts the autoencoder architecture but also assumes that the distribution of encoding features is similar to that of original features[30]. The framework of VAE is concisely described in Fig. 1 and precisely in Supplementary Fig. S1A. The structure of VAE is described in Supplementary File.

We obtained information on TF-related genes from the TRANSFAC database 7.0, which is publicly available at http://gene-regulation.com [31]. We selected "Factor", "homo sapiens", and "Organism Species (OS)" as the values of "option", "search term", and "table field to search in", respectively, yielding a list of 608 TF-related genes. Then, common genes between the TF-related genes and DEGs were used as input features of a classifier. Detailed information about the 608 genes and the TF-related genes overlapped with DEGs is presented in Supplementary Table S1.

To curate hub genes, we obtained gene-gene interaction (GGI) data, including 7,765 genes and 54,719 interactions from the Human Protein Reference Database, publicly available at http://www.hprd.org/ [32]. The process of extracting hub genes among DEGs consisted of several steps. We mapped DEGs obtained from each training set onto the data of GGI and calculated the number of edges of DEGs. To select a specific number of edges, we applied 10 different numbers of edges as thresholds ranging from 10 to 20 to ADNI, ANM1, and ANM2. Among the 10 thresholds, we selected a value of 10, which was similar to the number of CFG-based genes. Therefore, we defined genes with more than 10 edges as hub genes, which interact with more than 10 genes in the GGI database.

The CFG approach is a translational methodology that integrates multiple lines of external evidence from human and animal model studies[7,8]. There have been several studies using various CFG scoring methods, from which we selected two representative methods[6,33]. The first method[6] was to score genes validated by multi-genomic and experimental studies using the five criteria (Supplementary File). We assigned each DEG one point if the DEG satisfied one of the above five criteria, yielding a score ranging from 0 to 5 points by using a publicly accessible database at http://alzdata.org [6]. The second method is the database-based CFG scoring method that uses external lines of evidence[33]. We scored a gene as one point if the gene was included in the AD-related genes extracted from two databases, including Alzgene and DigSee[34,35], yielding a score ranging from 0 to 2. Bertram *et al.* constructed a publicly available, continuously updated database (AlzGene, http://www.alzgene.org) by performing systematic meta-analysis for each polymorphism with available genotype data in at least three case-control subjects[34]. The DigSee extracted gene-disease relationships by incorporating the text-mining method and the ML technique and included 4,494 disease types and 13,054 genes[35]. From the AlzGene and DigSee, we obtained a list of 680 and 1602 AD-related genes, respectively. Combining these two CFG scoring methods, we annotated all DEGs with numeric points ranging from 0 to 7. We defined DEGs with 3 or more points as highly informative AD genes. Lists of DEGs with CFG score (ADNI, ANM1, and ANM2) are presented in Supplementary Tables S2–S4.

Classifying methods. We utilized LR, L1-GLM, RF, SVM, and DNN as a classifying model. LR, developed by David Cox in 1958, is a standard method for binary classification[36]. L1-LR, first suggested by Tibshirani in 1996, is an extended version of LR applied by LASSO[37]. Due to a sparsity, the L1-LR can simultaneously perform two tasks: feature selection and classification[38,39]. The L1-LR needs λ, the tuning hyperparameter that controls the degree of the penalty, which is typically set as a value that gives the best performance via CV. However, the previous study showed that most weights were penalized to 0 after selecting λ by CV[20]. Therefore, we preliminarily applied 100 sequencing λ values ranging from 10^{-3} to 10^{-5}. The larger lambda was selected, the smaller number of genes were selected. Of 100 λ values, we selected the 50th λ value ($\lambda = 0.0001$) because this value is the elbow point, where the increasing tendency of the number of selected genes is reduced (Supplementary Fig. S2).

SVM, a robust ML method, is typically used to solve binary classification problems by finding a hyper-plane that maximizes the margin between two classes. Two hyperparameters are needed for the SVM algorithm: cost (C), which indicates the degree of penalty for misclassification, and gamma (γ), which defines the extent of the influence of a single training example. In this study, we adopted the Gaussian radial basis kernel for SVM. The "svm" function in the e1071 package was used to run the SVM algorithm, in which, by default, C and γ values were set to 1 and 1/(dimension of input features)[40], respectively. Note that another widely used package, Scikit-learn[41], also adopts default values of C and γ, similar to those of the e1071 package.

The RF algorithm, developed by Leo Breiman, utilizes an ensemble of classification trees, which include bootstrap samples and randomly selected variables[42,43]. Two hyperparameters are required in RF: the number of trees, ntree, and the number of randomly selected features, mtry. In this study, we determined ntree = 500, and ntry = $\sqrt{\text{data dimension}}$, which were default values.

DNN is a method of learning representative features with multi levels of representation, obtained via non-linear perceptrons, each of which transforms the representation at one level to that at a higher or abstract level with reduced dimension[44,45]. LeCun *et al.*[44] suggested that these higher levels of representation would amplify important aspects of the input for classification tasks and suppress irrelevant variations. In this study, a DNN architecture consists of two hidden layers, where the number of hidden nodes are $\left[\frac{N}{2}\right]$ and $\left[\frac{N}{4}\right]$ (N: number of input features) in the first and second hidden layers, respectively. We set the minimum of hidden nodes in the first and second hidden layers as 10 and 5, respectively. The hyperbolic tangent function between hidden layers and the sigmoid function in the final layer were employed. We hypothesized that more input features needed more iterations of training. Therefore, we determined the number of iteration as "N (number of input features) $\times$ 3", and also set the minimum and maximum number of iteration as 200 and 3000, respectively. We optimized our DNN model utilizing the Adagrad optimizer[46] when an input dimension was ≥ 800 or the AdamOptimizer when an input dimension was <800[47], and a learning rate was set to 0.001.

Results

Internal validation (CV within each dataset). For samples in ADNI, ANM1, and ANM2, the average chronological ages were 75.9, 74.1, and 76.7, and ratios of males were 49.7, 35.3, and 39.4%, respectively (Table 2). In ADNI, ages did not differ significantly between AD and CN samples, while ages in ANM1 and ANM2 showed a significant difference between AD and CN samples. The gender difference between AD and CN samples was not significant in all datasets (Table 2).

In the internal validation (five-fold CV, ADNI), DEGs ranging from 72 to 922 were curated from each training set, and are presented in Supplementary Table S5. Similarly, DEGs ranging from 850 to 1617 and those from 187 to 790 were selected in each five-fold CV for ANM1 and ANM2, respectively (Supplementary Table S5). The numbers of DEGs of ADNI varied most with a standard deviation (SD) of 352.7 while those of ANM2 were most consistent (an SD of 254.4). When all DEGs were used as input features in ADNI, DNN and RF outperformed the other methods (Fig. 2A). In ANM1, L1-LR, SVM, RF, and DNN showed AUC values greater than 0.80 (Fig. 2B). In ANM2, SVM was the best performing classifier for distinguishing AD when all DEGs were used as input features (Fig. 2C).

	ADNI			ANM1			ANM2		
	CN	AD	P value	CN	AD	P value	CN	AD	P value
Number of samples, n	136	63		104	145		134	139	
Age, years	75.62 ± 6.76	76.51 ± 7.63	0.43	72.38 ± 6.34	75.40 ± 6.58	<0.001	75.29 ± 6.02	77.89 ± 6.67	<0.001
Gender (male), n	64 (47.1)	35 (55.6)	0.265	42 (40.4)	46 (31.7)	0.159	53 (39.6)	54 (38.8)	0.762

Table 2. Demographic characteristics in ADNI, ANM1, and ANM2. ADNI: Alzheimer's Disease Neuroimaging Initiative; ANM: AddNeuroMed; CN: healthy control; AD: Alzheimer's disease; Continuous variables are presented as mean ± standard deviation, and categorical variables are as number (percent, %).

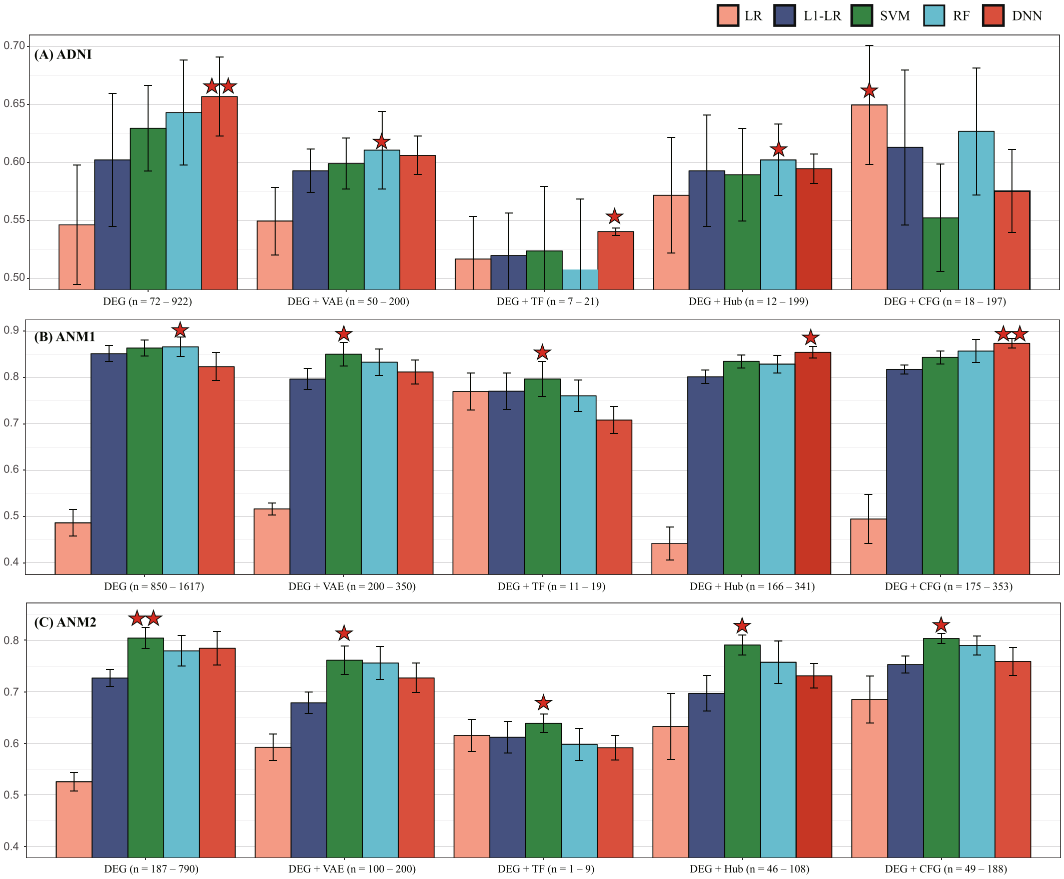


Figure 2. Internal validation. (**A–C**) illustrate performance (y-axis: AUC) of classifying AD from CN in ADNI, ANM1, and ANM2, respectively. X-axes of (**A–C**) describe feature selection methods, including DEG, VAE, TF, and CFG. Different colors of columns indicate classifiers, including LR, L1-LR, SVM, and DNN. One star denotes a model with the best performance among five classifiers, and two stars denote the best model among five classifiers and five feature selection methods. The ranges of numbers indicate the numbers of selected genes in each feature selection method, and all numbers of selected genes are described in the Supplementary Table S5. AUC, area under the curve; AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; ANM, AddNeuroMed; DEG, differentially expressed gene; VAE, variational autoencoder; TF, transcription factor; LR, logistic regression; L1-LR, L1-regularized LR; SVM, support vector machine; RF, random forest; DNN, deep neural network.

In ADNI, ANM1 and ANM2, when the VAE method was used for dimension reduction, the performances were improved only for the LR classifier compared with the DEGs. For the other methods, the application of VAE may lose important information (Fig. 2, Supplementary Table S5).

We extracted TF-related genes ranging from zero to 21 among DEGs of ADNI (Supplementary Table S5). In the same way, 11 to 19 TF-related genes and zero to 9 TF-related genes were selected for ANM1 and ANM2, respectively. In ADNI, ANM1, and ANM2, the mean AUC values of all classifying methods, except for LR, decreased when a set of TF-related genes were used as input features (Fig. 2).

We selected 12 to 199, 166 to 341, and 46 to 108 hub genes with more than 10 edges in the GGI network among DEGs in ADNI, ANM1, and ANM2, respectively (Supplementary Table S5). As a result, the majority of classifying methods did not show improved performance, except for LR, DNN, and LR in ADNI, ANM1, and ANM2, respectively (Fig. 2).

The CFG method was utilized to select 18 to 197 genes from DEGs in ADNI, showing that LR and L1-LR showed improved performance. In ANM1 and ANM2, after the CFG scoring was used, 850 to 1617 and 187 to 790 of DEGs were shrunk to 175 to 353 and 49 to 188 genes, respectively. Two classifying methods (LR and DNN) in ANM1 and three (LR, L1-LR, and RF) in ANM2 were better than those using all DEGs as input, respectively (Fig. 2).

For each feature selection method, most comparisons across the five classifiers did not yield significant differences in performance (measured by a paired *t*-test, Supplementary File) in ADNI, ANM1, and ANM2 (Supplementary Table S5), except that LR showed statistically lower performances compared to other methods in ANM1 and ANM2.

Among five feature selection approaches, on average, the DEG provided the best performance in ADNI ($p = 0.109$ measured by a *t*-test, please refer to the Supplementary File) and ANM1 ($p = 0.631$), and the "DEG + CFG" ($p = 0.002$) showed the best performance in ANM2 (Fig. 2).

External validation (Cross datasets analysis). When 334 DEGs extracted from ADNI served as input features for predicting AD patients in ANM1, the best and second-best performing classifiers were L1-LR and SVM with 0.70 and 0.66 AUC values, respectively. When ADNI and ANM2 were used as training and test datasets with input features of the 334 DEGs, respectively, L1-LR manifested the best AUC (0.69) among classifiers. Note that the process of extracting 334 DEGs in ADNI was completely independent of ANM1 and ANM2 datasets.

When using VAE (334 to 100) and TF-related genes (seven genes), the performances were not better than that of the method using 334 DEGs were proposed when arranging ADNI and ANM (ANM1, ANM2) as training and testing datasets, respectively (Fig. 3A,B, Supplementary Table S6).

We tested the ANM1 dataset with 67 hub genes selected from ADNI and achieved lower AUC than DEGs. However, the best performance (AUC: 0.76) was acquired when testing ANM2 with the 67 hub genes from ADNI and SVM (Fig. 3B). Using 81 DEGs with CFG ≥ 3 as input features, most classifiers did not show improved performances, except for LR (ANM1) and RF (ANM2) (Fig. 3A,B).

The best performance in ANM1 (testing set 1) was an AUC of 0.70 when the 334 DEGs and L1-LR served as the feature selection method and classifying model, respectively (Fig. 3A). In the case of ANM2 (testing set 2), an AUC of 0.76 was the best result achieved using the hub method and SVM (Fig. 3B).

When 1604 DEGs from ANM1 were used for training, SVM showed the best performance with AUCs of 0.62 and 0.79 for test sets ADNI and ANM2, respectively (Fig. 3C,D). When we used VAE (1604 to 300), 18 TF-related genes, 331 hub genes, and the CFG method, performances of most classifying methods decreased. The best performances in ADNI and ANM2 occurred when DEGs and SVM were adopted as the input features and the classifier, respectively.

When 697 DEGs from ANM2 were used for training, DNN showed the best performances of 0.65 and 0.85 for both ADNI and ANM1, respectively (Fig. 3E,F). Using reduced 200 dimensions from 697 DEGs via VAE and TF-related genes as input, overall performance decreased compared with using 697 DEGs. When testing ADNI, the hub gene offered the best performance (AUC: 0.66, DNN); however, the CFG performed best when testing ANM2 (AUC: 0.86, DNN).

In the comparisons across the five classifiers, the performances of four classifiers (L1-LR, SVM, RF, and DNN) were enhanced as compared with that of LR. Among the four classifiers, most comparative analyses showed an insignificant difference in terms of performances (*p*-values were measured by Venkatraman's method, refer to the Supplementary File and Supplementary Table S6).

Among five feature selection methods, on average, the DEG ($p = 0.146$ measured by a *t*-test, Supplementary File) provided the best performance compared to other methods when ADNI and ANM1 were arranged as training- and test-sets, the "DEG + Hub genes" (*t*-test, $p = 0.2$) showed the best performance between ADNI (training) and ANM2 (test), and the "DEG + CFG" (*t*-test, $p = 0.039$) yielded the best result between ANM1 (training) and ANM2 (test).

Pathway analysis of informative genes in each dataset. We found that the performances of discriminating AD were best when "DEG" or "DEG + CFG" were used as feature selection methods. Therefore, we performed pathway analysis of DEGs (334 genes of ADNI, 1604 of ANM1, and 697 of ANM2) and "DEG + CFG" (81 genes of ADNI, 334 of ANM1, and 169 of ANM2) using the KEGG pathways[48] and Gene Ontology[49] obtained from Molecular Signature Database (MSigDB)[50]. We removed general pathways that consist of ≥ 500 genes. A pathway enrichment test was performed using a hypergeometric test followed by multiple comparison correction (Benjamini–Hochberg method) and pathways with q-values <0.05 were considered significantly enriched. As a result, no pathways (ADNI: DEG), 26 (ANM1: DEG), 57 (ANM2: DEG), 7 (ADNI: DEG + CFG), 340 (ANM1: DEG + CFG), and 366 pathways (ANM2: DEG + CFG) were selected (Supplementary Table S7).

In ADNI, the representative pathways enriched via 81 genes (DEG + CFG) were as follows: immune system process, ErbB signaling, and lipopolysaccharide mediated signaling pathways (Fig. 4).

When using only DEG in ANM1 and ANM2, 26 common pathways were obtained, and the following pathways were notable: mitochondrial translation, oxidative phosphorylation, and ribonucleoprotein complex biogenesis (Fig. 4). DEGs with CFG in ANM1 and ANM2 were commonly enriched in 51 pathways, such as negative regulation of canonical Wnt signaling pathway, activation of the immune response, and myeloid cell homeostasis (Fig. 4).

Common pathways among four lists of pathways (two feature selection methods and two datasets [ANM1, ANM2]) were 17, and representative pathways are as follows: ribosome, the establishment of protein localization

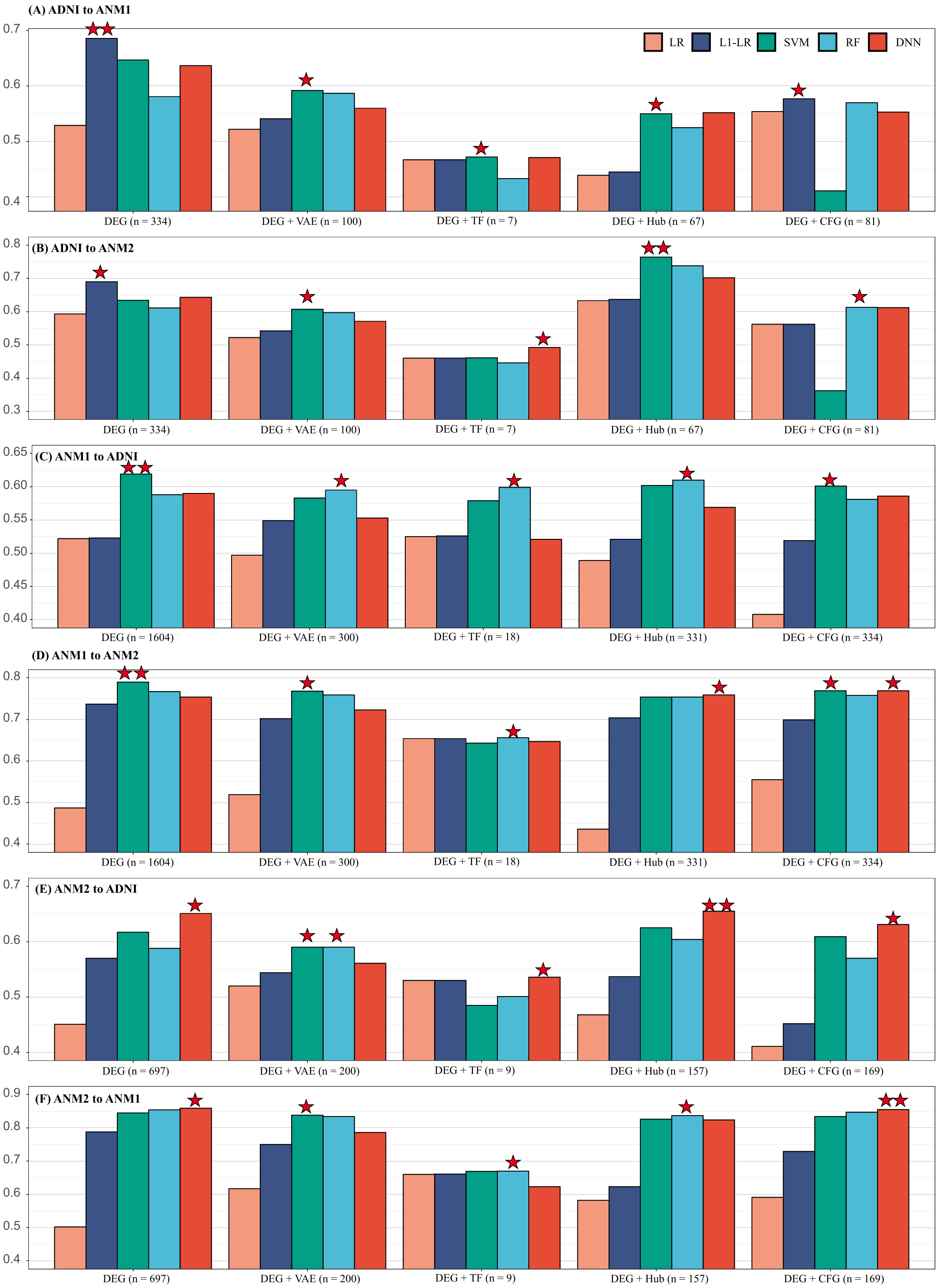
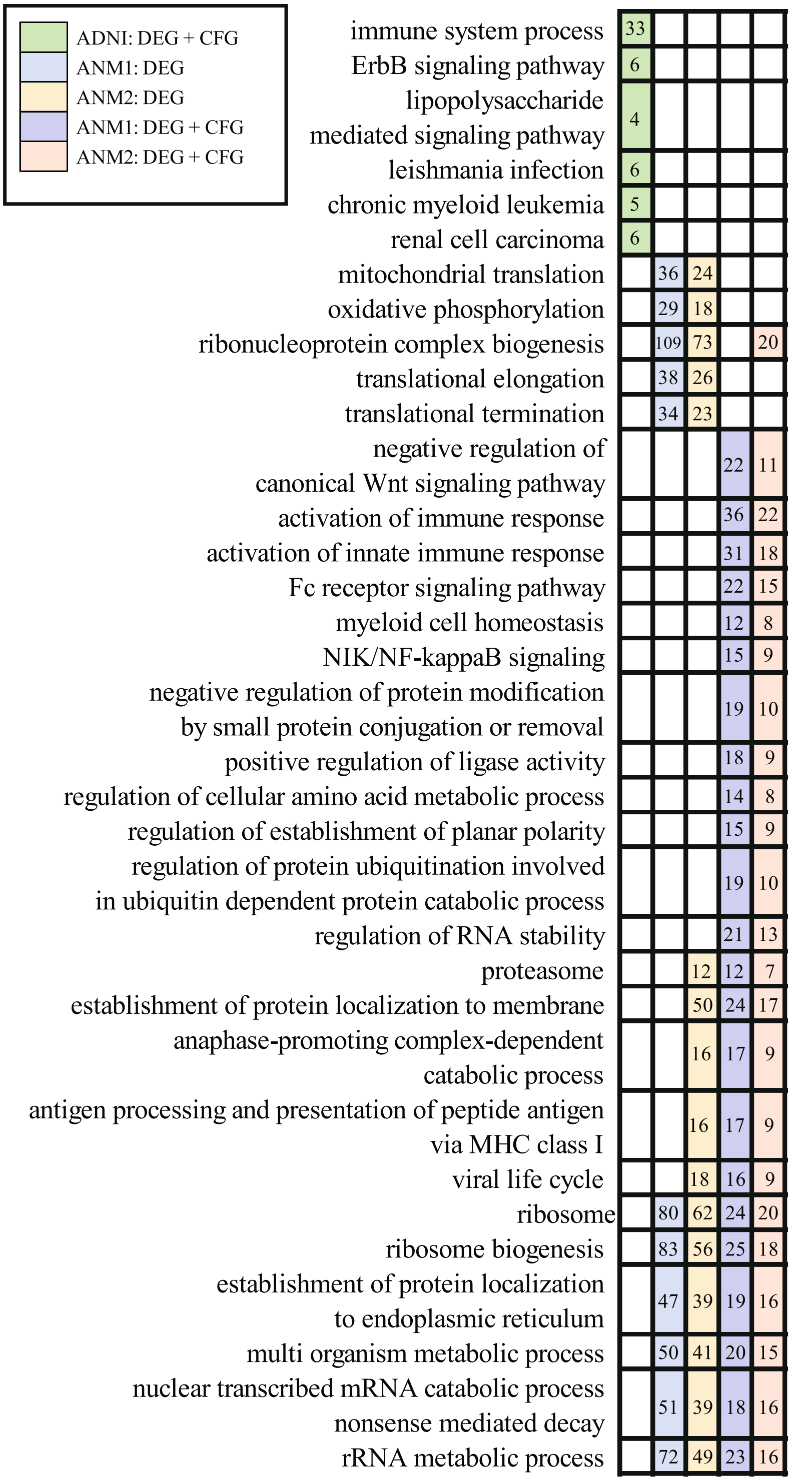


Figure 3. External validation. (**A**) ADNI and ANM1 (**B**) ADNI and ANM2 (**C**) ANM1 and ADNI (**D**) ANM1 and ANM2 (**E**) ANM2 and ADNI (**F**) ANM2 and ANM1 are arranged as training and testing dataset, respectively. The y-axes in each graph show performance measured by AUC. One star denotes a model with the best performance among five classifiers, and two stars denote the best model among five classifiers and five feature selection methods. The number indicates the number of selected genes in each feature selection method. AUC, area under the curve; AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; ANM, AddNeuroMed.

to the endoplasmic reticulum, and multi organism metabolic process (Fig. 4). Collectively, highly informative genes in ADNI were enriched in immune and inflammatory pathways. Differentially expressed genes from ANM were associated with energy metabolism (mitochondria and oxidative phosphorylation), and "DEG + CFG" from ANM showed significant enrichment for Wnt-related and immune pathways (Fig. 4).

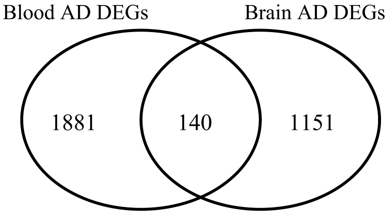


🟩 ADNI: DEG + CFG
🟦 ANM1: DEG
🟨 ANM2: DEG
🟪 ANM1: DEG + CFG
🟥 ANM2: DEG + CFG

immune system process	33				
ErbB signaling pathway	6				
lipopolysaccharide mediated signaling pathway	4				
leishmania infection	6				
chronic myeloid leukemia	5				
renal cell carcinoma	6				
mitochondrial translation		36	24		
oxidative phosphorylation		29	18		
ribonucleoprotein complex biogenesis		109	73		20
translational elongation		38	26		
translational termination		34	23		
negative regulation of canonical Wnt signaling pathway				22	11
activation of immune response				36	22
activation of innate immune response				31	18
Fc receptor signaling pathway				22	15
myeloid cell homeostasis				12	8
NIK/NF-kappaB signaling				15	9
negative regulation of protein modification by small protein conjugation or removal				19	10
positive regulation of ligase activity				18	9
regulation of cellular amino acid metabolic process				14	8
regulation of establishment of planar polarity				15	9
regulation of protein ubiquitination involved in ubiquitin dependent protein catabolic process				19	10
regulation of RNA stability				21	13
proteasome			12	12	7
establishment of protein localization to membrane			50	24	17
anaphase-promoting complex-dependent catabolic process			16	17	9
antigen processing and presentation of peptide antigen via MHC class I			16	17	9
viral life cycle			18	16	9
ribosome		80	62	24	20
ribosome biogenesis		83	56	25	18
establishment of protein localization to endoplasmic reticulum		47	39	19	16
multi organism metabolic process		50	41	20	15
nuclear transcribed mRNA catabolic process nonsense mediated decay		51	39	18	16
rRNA metabolic process		72	49	23	16

Figure 4. Significant pathways in ADNI, ANM1, and ANM2. Right-side matrix is marked with the color assigned to each data if the genes of each data are enriched in left-side pathways. Numbers in each square indicate selected genes among different pathways. ADNI, Alzheimer's Disease Neuroimaging Initiative; ANM, AddNeuroMed.

Common bio-signature between blood and brain AD samples. We analyzed the brain gene expression datasets (GEO: GSE33000), including 157 AD samples and 310 CN. We curated 1291 DEGs using the "lmFit" function in the limma package followed by a multiple comparison correction and a fold change (FC) (FDR <0.05 and $|\log_2(\mathrm{FC})| > 0.2$). When comparing 2021 blood DEGs (union of DEGs among ADNI, ANM1, and ANM2), 140 of 1291 brain DEGs were common with blood DEGs, which were enriched with two KEGG



Pathways enriched by the 140 common genes	# of Overlapped Genes
Hematopoietic cell lineage	7
Leishmaniasis	7
immune response	36
inflammatory response	26
MyD88-dependent toll-like receptor signaling pathway	5
respiratory burst	4
toll-like receptor 4 signaling pathway	4

Figure 5. Common AD bio-signatures between blood and brain. AD, Alzheimer's Disease; DEG, differentially expressed gene.

and 31 GO pathways (Fig. 5). Representative pathways included immune response, inflammatory response, MyD88-dependent toll-like receptor signaling, and toll-like receptor 4 signaling pathways (Fig. 5).

Discussion

In this study, we identified AD-related genes by means of DEGs, the TF database, connectivity in the gene network, and CFG. Considering both internal and external validations, this study showed that all DEGs and the DEGs with CFG could accurately identify AD. Among the previous studies listed in Table 1, a study by Li *et al.*[20] classified AD patients most accurately in the ANM dataset. However, in their study[20], it seems that both ANM1 and ANM2 datasets were used for selecting features (six genes). After curating the six AD-related genes, they used them as input features for the SVM model, yielding AUCs of 0.86 and 0.873 when testing ANM1 and ANM2, respectively[20]. With the same six genes and the same training- and test-set described by Li *et al.*[20], we measured prediction performances and obtained results that were consistent with those of Li *et al.*[20]. Afterward, we predicted an independent data ADNI using these six genes as input features to the SVM method, and obtained AUCs 0.62 (training dataset: ANM1) and 0.57 (ANM2). This method[20] was not as effective as our method using DEGs as input features to the same method, which had AUCs of 0.62 (training dataset: ANM1) and 0.63 (ANM2).

The primary objective of this study was to investigate the prediction of AD patients using AD-related genes obtained across different datasets. The performances in the external validation were high between ANM1 and ANM2, consistent with the previous studies[18–20]. However, the prediction accuracies were low between ADNI and ANM (ANM1 and ANM2), and there was no study using these datasets as external validation. Several studies have suggested the limitations of ADNI. First, gene expression differences between AD and CN in ADNI were low compared to those in ANM, yielding difficulty in extracting DEGs in ADNI. Li *et al.*[20] faced a similar problem, and selected DEGs based on nominal p-values < 0.05 because no gene passed a multiple testing correction. To overcome this limitation, we removed probes with low intensity by a previously validated method[25]. Furthermore, we attempted several methods to curate DEGs, including a t-test, limma[29], and SAM[28], and found that SAM, which is based on the permutation, could extract DEGs (FDR < 0.05). Second, we observed that for ADNI, the numbers of DEGs varied most across five CV sets among ADNI, ANM1, and ANM2 (the highest SDs for the numbers of DEGs were as follows, ADNI: 352.7, ANM1: 309.3, ANM2: 254.4). This observation might partially explain why in the internal validation, the AUCs for ADNI were lower than those for ANM1 and ANM2. Third, the qualities of gene expression data for some samples in ADNI were low. When all samples in ANDI were used, performances in both internal and external validations were low. The AUC values of the highest performance classifiers for internal and external validation were only 0.613 (Internal validation, ADNI), 0.601 (External validation, ANM1 to ADNI), and 0.63 (External validation, ANM2 to ADNI), respectively. Thus, in this study, we used gene expression samples with RIN values ≥ 6.9, resulting in increased performances in Figs. 2 and 3.

In the external validation, L1-LR, SVM, and DNN showed the best performance. In a previous study, L1-LR and SVM performed well as a feature selection method and a classifier for AD classification, respectively[20]. We observed that DNN did not outperform other classifiers in all cases, which might be because the relatively small size of samples was composed of high dimensional information (i.e., the number of genes) and the number of samples was insufficient to learn all perceptrons of DNN. Furthermore, when we applied several settings of DNN to improve performance, including drop-out and early stopping, we found that although the maximum AUC value increased, the range of the AUC values become more varied than that of the default setting. In future work, the performance will be improved by adjusting parameters and modifying the architecture of DNN when more data are available[51,52].

We investigated DEG genes with high CFG scores for their correlation to AD. Among 334 DEGs in the ADNI dataset, four genes (TGFB1, RAB11A, MAPK3, and RTN4) showed a maximum CFG score of five points. TGFB1 reportedly increases the risk of developing late-onset AD[53], ERK (MAPK3) is expectedly activated in AD brains and involved in tau phosphorylation and amyloid deposition[54]. According to recent reports, two additional genes (RAB11A, RTN4) are also related to AD[55,56]. Among 1604 DEGs in the ANM1 dataset, five genes (TIMP1, CD14, FADD, CAMK2G, and FCER1G) had the highest CFG score of six points. MMP-9-TIMP1 pathway was known to be stimulated by Abeta 25–35 fragment to eliminate amyloid deposition from AD brains[57]. The lipopolysaccharide (LPS) receptor CD14 also reportedly contributes to neuroinflammation in AD[58]. FCER1G is one of the microglial-specific genes, and the microglial is considered a major causal factor in AD[59]. Among 697 DEGs in the ANM2 dataset, ten genes (DBI, CDK5R1, SORL1, CTNNA1, CTSS, CAPN1, NFKBIA, SERPINA1, CST3, and VIM) had maximum scores of five CFG points. CDK5R1 is known to be closely related to AD onset and progression[60], SORL1 interacts with the movement of APP and plays a possible role in AD progression[61], whereas CTNNA1 is critical to the folding and lamination of the cerebral cortex and is involved in AD pathogenesis[62].

We observed several enriched pathways with AD-related genes in blood samples in ADNI and ANM, which were also enriched in human or mouse brain tissues. The main pathways of ADNI were related to immune response. A study that found conserved genetic signals in mice and human brain tissues reported that genes related to early and late-stage AD were significantly enriched in immune system processes[63]. Additionally, infection-related pathways (lipopolysaccharide mediated signaling pathway) were enriched in ADNI, which was activated in the human brain with AD[64], and associated with the increased risk of AD[65]. The DEGs in ANM were significantly enriched with mitochondria-related pathways. Liang *et al.*[13] demonstrated that AD cases had significantly down-regulated expression of the nuclear genes encoding subunits of the mitochondrial electron transport chain in several brain regions. The DEGs with CFG ≥ 3 in ANM were enriched with the Wnt signaling pathway, which is associated with the developmental process of the nervous system, and especially its association with synaptogenesis was validated in a mouse brain model[66] and the protective role of neurodegeneration in AD rat model[67].

In Table 2, ANM1 and ANM2 showed a significant difference in terms of age between AD and CN. Although gender difference did not significantly differ between AD and CN, other studies reported gender difference in the AD risk[68]. Thus, we adjusted for age as well as gender for the ADNI, ANM1, and ANM2 datasets. Afterward, we measured AD predictive performances using five feature selection methods and five classifiers. The detailed adjustment procedure is described in the Supplementary File. As a result, we observed that classifying performances did not significantly differ after the adjustment (Supplementary Fig. S3).

There are three subtypes of AD, including preclinical AD[69], prodromal AD[70], and AD dementia. The subtypes of AD are diagnosed using the amyloid positron emission tomography (PET) scan. However, the ANM datasets (ANM1 and ANM2) did not include amyloid PET results. While the ADNI dataset had amyloid PET data, samples with gene expressions were not explicitly classified by the amyloid PET image. Thus, instead of analyzing the three subtypes of ADs, we analyzed MCI samples in the ADNI, ANM1, and ANM2 diagnosed by the NINCDS-ADRDA criteria[2]. First, we compared FCs between two pairs of datasets, AD *vs*. CN and MCI *vs*. CN, for each dataset (Supplementary Fig. S4). In detail, we compared $\log_2$FC for all genes ($n = 8{,}835$), DEGs ($n = 334$, 1604 and 697 for ADNI, ANM1, and ANM2, respectively), and DEG with CFG ($n = 81$, 334, and 169 for ADNI, ANM1, and ANM2, respectively) using Spearman correlation. As a result, we observed positive correlations between $\log_2$ (AD/CN) and $\log_2$ (MCI/CN) in all datasets and several gene sets (all genes, DEG, and DEG with CFG) (Supplementary Fig. S4), suggesting that genes are similarly upregulated or downregulated in AD and MCI.

We further integrated the ADNI, ANM1, and ANM2 datasets, and investigated the AD prediction performance on the integrated dataset (Supplementary File). To integrate the three datasets and select features from it, we applied four different approaches. First, the ComBat method, which was mainly used for internal and external validations, was used to remove batch effects among these datasets[27], and then DEGs were selected using the "lmFit" function in the limma package. Second, the scaling & quartiling method by Mohammadi-Dehcheshmeh *et al.*[71] was used to remove batch effects further, and then DEGs were selected using the "lmFit" function. Third, DEGs were computed for each dataset via the moderate *t*-test, and then rankings of *p*-values were used to curate meta-DEGs among three datasets[72]. Fourth, DEGs were computed for each dataset via the moderate *t*-test, and then *p*-values were combined via Fisher's method[73]. The gene expression values normalized by the ComBat method showed significant correlations with those yielded by the other three approaches (Supplementary Fig. S5). Furthermore, the DEGs curated by the first ComBat approach significantly correlated with those by the other three methods (Supplementary Fig. S6). We evaluated the AD predictive performance by the three-fold CV. When five classifiers, LR, L1-LR, SVM, RF, and DNN, were used for training and test, the ComBat approach obtained AUC values of 0.5, 0.7, 0.8, 0.79, and 0.79, respectively (Supplementary Fig. S7). This result shows that these three datasets can be integrated for classifying AD and CN. Also, the ComBat method outperformed the other three methods in terms of the average AUC values of the five classifiers (Supplementary Fig. S7).

Most studies have used the NINCDS-ADRDA criteria made in 1984[2] for classifying subjects as AD and CN samples. The NINCDS-ADRDA criteria are symptom- and individual doctor-based diagnosis, and therefore, may yield inconsistencies between different datasets. Using a single nucleotide polymorphism dataset in ADNI, Apostolova *et al.*[74] curated seven variants using brain amyloidosis as a dependent variable, while only one variant (FERMT2) was found using the AD stage determined by the NINCDS-ADRDA as a dependent variable. In addition, Edmonds *et al.*[75] suggested that some MCI samples diagnosed by the NINCDS-ADRDA are false positives. In 2007, a revised version of the NINCDS-ADRDA criteria, which more focused on the pathology of AD rather than clinical symptoms, was introduced[3]. In the future, if genomic data composing participants determined by the revised diagnostic criteria of AD are available, more AD-related genes and pathways can be identified.

Conclusions

In this study, we showed that expression values of AD-related genes obtained from blood samples of ADNI, ANM1 and ANM2 could classify AD and CN. Additionally, we observed that AD-related genes from blood samples were enriched with several pathways including immune, inflammation, energy metabolism, and Wnt signaling, which are consistent with observations from brain tissue-based studies. Collectively, AD-related genes from blood samples contribute to the development of blood-based AD diagnostic and treatment tools.

Data availability

ADNI and ANM datasets are publicly available (ADNI, http://adni.loni.usc.edu/; ANM, https://www.ncbi.nlm.nih.gov/geo/).

Received: 2 October 2019; Accepted: 11 February 2020;
Published online: 26 February 2020

References

1. Hebert, L. E., Weuve, J., Scherr, P. A. & Evans, D. A. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurol.* **80**, 1778–1783 (2013).
2. McKhann, G. *et al.* Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurol.* **34**, 939–939 (1984).
3. Dubois, B. *et al.* Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. *Lancet Neurol.* **6**, 734–746 (2007).
4. Lambert, J.-C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* **41**, 1094 (2009).
5. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452 (2013).
6. Xu, M. *et al.* A systematic integrated analysis of brain expression profiles reveals YAP1 and other prioritized hub genes as important upstream regulators in Alzheimer's disease. *Alzheimers Dement.* **14**, 215–229 (2018).
7. Niculescu, A. B. & Le-Niculescu, H. Convergent Functional Genomics: what we have learned and can learn about genes, pathways, and mechanisms. *Neuropsychopharmacology* **35**, 355 (2010).
8. NICULESCU, A. B. III *et al.* Identifying a series of candidate genes for mania and psychosis: a convergent functional genomics approach. *Physiol. Genomics* **4**, 83–91 (2000).
9. Mueller, S. G. *et al.* Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement.* **1**, 55–66 (2005).
10. Lovestone, S. *et al.* AddNeuroMed—the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Ann. N. Y. Acad. Sci.* **1180**, 36–46 (2009).
11. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.* **39**, D1005–D1010 (2010).
12. Durrenberger, P. F. *et al.* Common mechanisms in neurodegeneration and neuroinflammation: a BrainNet Europe gene expression microarray study. *J. Neural Transm.* **122**, 1055–1068 (2015).
13. Liang, W. S. *et al.* Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proc. Natl Acad. Sci. USA* **105**, 4441–4446 (2008).
14. Cooper, Y. A. *et al.* Progranulin levels in blood in Alzheimer's disease and mild cognitive impairment. *Ann. Clin. Transl. Neurol.* **5**, 616–629 (2018).
15. Booij, B. B. *et al.* A gene expression pattern in blood for the early detection of Alzheimer's disease. *J. Alzheimers Dis.* **23**, 109–119 (2011).
16. Lunnon, K. *et al.* A blood gene expression marker of early Alzheimer's disease. *J. Alzheimers Dis.* **33**, 737–753 (2013).
17. Sood, S. *et al.* A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. *Genome Biol.* **16**, 185 (2015).
18. Voyle, N. *et al.* A pathway based classification method for analyzing gene expression for Alzheimer's disease diagnosis. *J. Alzheimers Dis.* **49**, 659–669 (2016).
19. Li, H. *et al.* Identification of molecular alterations in leukocytes from gene expression profiles of peripheral whole blood of Alzheimer's disease. *Sci. Rep.* **7**, 14027 (2017).
20. Li, X. *et al.* Systematic analysis and biomarker study for Alzheimer's disease. *Sci. Rep.* **8**, 17394 (2018).
21. Team RC. R: A language and environment for statistical computing (2013).
22. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
23. Abadi M. *et al.* Tensorflow: A system for large-scale machine learning. In: *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (ed^(eds) (2016).
24. Hokama, M. *et al.* Altered expression of diabetes-related genes in Alzheimer's disease brains: the Hisayama study. *Cereb. Cortex* **24**, 2476–2488 (2013).
25. Antonell, A. *et al.* A preliminary study of the whole-genome expression profile of sporadic and monogenic early-onset Alzheimer's disease. *Neurobiol. Aging* **34**, 1772–1778 (2013).
26. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
27. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinforma.* **28**, 882–883 (2012).
28. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
29. Smyth, G. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article 3 (2004).
30. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:13126114* (2013).
31. Matys, V. *et al.* TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
32. Keshava Prasad, T. *et al.* Human protein reference database—2009 update. *Nucleic Acids Res.* **37**, D767–D772 (2008).
33. Breen, M. S. *et al.* Candidate gene networks and blood biomarkers of methamphetamine-associated psychosis: an integrative RNA-sequencing report. *Transl. Psychiatry* **6**, e802 (2016).
34. Bertram, L., McQueen, M. B., Mullin, K., Blacker, D. & Tanzi, R. E. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.* **39**, 17–23 (2007).
35. Kim, J., Kim, J. J. & Lee, H. An analysis of disease-gene relationship from Medline abstracts by DigSee. *Sci. Rep.* **7**, 40154 (2017).
36. Cox, D. R. The regression analysis of binary sequences [with discusion]. *J. R. Stat. Soc. Ser. B Stat Methodol.* **20**, 215–242 (1958).
37. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat Methodol.* **58**, 267–288 (1996).
38. Ayers, K. L. & Cordell, H. J. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.* **34**, 879–891 (2010).
39. Algamal, Z. Y. & Lee, M. H. Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert. Syst. Appl.* **42**, 9326–9332 (2015).
40. Dimitriadou, E. *et al.* Package 'e1071'. *R Software package, avaliable at*, http://cran rproject org/web/packages/e1071/index html (2009).
41. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
42. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
43. Ebrahimi, M., Lakizadeh, A., Agha-Golzadeh, P., Ebrahimie, E. & Ebrahimi, M. Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: a new vista in engineering enzymes. *PLoS One* **6**, e23146 (2011).
44. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nat.* **521**, 436 (2015).
45. Jamali, A. A. *et al.* DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. *Drug. discovery today* **21**, 718–724 (2016).
46. Duchi, J., Hazan, E. & Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**, 2121–2159 (2011).
47. Kingma, D. P. & Ba, J. A. A method for stochastic optimization. arXiv 2014. *arXiv preprint arXiv:14126980*, (2019).

48. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
49. Consortium, G. O. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
50. Liberzon, A. *et al*. Molecular signatures database (MSigDB) 3.0. *Bioinforma.* **27**, 1739–1740 (2011).
51. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
52. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (ed) (2010).
53. Bosco, P. *et al*. Role of the Transforming-Growth-Factor-beta1 Gene in Late-Onset Alzheimer's Disease: Implications for the Treatment. *Curr. genomics* **14**, 147–156 (2013).
54. Zhu, X., Lee, H. G., Raina, A. K., Perry, G. & Smith, M. A. The role of mitogen-activated protein kinase pathways in Alzheimer's disease. *Neurosignals* **11**, 270–281 (2002).
55. Roy, J., Sarkar, A., Parida, S., Ghosh, Z. & Mallick, B. Small RNA sequencing revealed dysregulated piRNAs in Alzheimer's disease and their probable role in pathogenesis. *Mol. Biosyst.* **13**, 565–576 (2017).
56. Masliah, E. *et al*. Genetic deletion of Nogo/Rtn4 ameliorates behavioral and neuropathological outcomes in amyloid precursor protein transgenic mice. *Neurosci.* **169**, 488–494 (2010).
57. Hernandez-Guillamon, M. *et al*. Neuronal TIMP-1 release accompanies astrocytic MMP-9 secretion and enhances astrocyte proliferation induced by beta-amyloid 25-35 fragment. *J. Neurosci. Res.* **87**, 2115–2125 (2009).
58. Liu, Y. *et al*. LPS receptor (CD14): a receptor for phagocytosis of Alzheimer's amyloid peptide. *Brain* **128**, 1778–1789 (2005).
59. Efthymiou, A. G. & Goate, A. M. Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk. *Mol. Neurodegener.* **12**, 43 (2017).
60. Spreafico, M., Grillo, B., Rusconi, F., Battaglioli, E. & Venturin, M. Multiple Layers of CDK5R1 Regulation in Alzheimer's Disease Implicate Long Non-Coding RNAs. *Int J Mol Sci* **19** (2018).
61. Yin, R. H., Yu, J. T. & Tan, L. The Role of SORL1 in Alzheimer's Disease. *Mol. Neurobiol.* **51**, 909–918 (2015).
62. Smith, A., Bourdeau, I., Wang, J. & Bondy, C. A. Expression of Catenin family members CTNNA1, CTNNA2, CTNNB1 and JUP in the primate prefrontal cortex and hippocampus. *Brain Res. Mol. Brain Res* **135**, 225–231 (2005).
63. Gjoneska, E. *et al*. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nat.* **518**, 365 (2015).
64. Zhan, X. *et al*. Gram-negative bacterial molecules associate with Alzheimer disease pathology. *Neurol.* **87**, 2324–2332 (2016).
65. Zhan, X., Stamova, B. & Sharp, F. R. Lipopolysaccharide associates with amyloid plaques, neurons and oligodendrocytes in Alzheimer's disease brain: a review. *Front. Aging Neurosci.* **10**, 42 (2018).
66. Ciani, L. *et al*. Wnt7a signaling promotes dendritic spine growth and synaptic strength through Ca2+/Calmodulin-dependent protein kinase II. *Proc. Natl Acad. Sci. USA* **108**, 10732–10737 (2011).
67. De Ferrari, G. V. *et al*. Activation of Wnt signaling rescues neurodegeneration and behavioral impairments induced by beta-amyloid fibrils. *Mol. Psychiatry* **8**, 195–208 (2003).
68. Vina, J. & Lloret, A. Why women have more Alzheimer's disease than men: gender and mitochondrial toxicity of amyloid-beta peptide. *J. Alzheimers Dis.* **20**(Suppl 2), S527–533 (2010).
69. Sperling, R. A. *et al*. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* **7**, 280–292 (2011).
70. Albert, M. S. *et al*. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* **7**, 270–279 (2011).
71. Mohammadi-Dehcheshmeh, M. *et al*. Unified Transcriptomic Signature of Arbuscular Mycorrhiza Colonization in Roots of Medicago truncatula by Integration of Machine Learning, Promoter Analysis, and Direct Merging Meta-Analysis. *Front. plant. Sci.* **9**, 1550 (2018).
72. Sharifi, S. *et al*. Integration of machine learning and meta-analysis identifies the transcriptomic bio-signature of mastitis disease in cattle. *PLoS One* **13**, e0191227 (2018).
73. Farhadian, M., Rafat, S. A., Hasanpur, K., Ebrahimi, M. & Ebrahimie, E. Cross-Species Meta-Analysis of Transcriptomic Data in Combination With Supervised Machine Learning Models Identifies the Common Gene Signature of Lactation Process. *Front. Genet.* **9**, 235 (2018).
74. Apostolova, L. G. *et al*. Associations of the top 20 Alzheimer disease risk variants with brain amyloidosis. *JAMA Neurol.* **75**, 328–341 (2018).
75. Edmonds, E. C. *et al*. Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors. *Alzheimers Dement.* **11**, 415–424 (2015).

Acknowledgements

This work has been supported by the Bio & Medical Technology Development Program of NRF funded by the Korean government (MSIT) (NRF-2018M3C7A1054935) and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant Number: HI18C0460).

Author contributions

H.L. contributed to the study concept and design. T.L. took part in the acquisition of the data and the machine learning algorithms. H.L. and T analyzed and interpreted results. T.L. and H.L. wrote the manuscript. H.L. took part in the study supervision and coordination. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-60595-1.

Correspondence and requests for materials should be addressed to H.L.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.