

UCSF

UC San Francisco Previously Published Works

Title

GenEpi: gene-based epistasis discovery using machine learning.

Permalink

<https://escholarship.org/uc/item/9tm364wd>

Journal

BMC bioinformatics, 21(1)

ISSN

1471-2105

Authors

Chang, Yu-Chuan

Wu, June-Tai

Hong, Ming-Yi

et al.

Publication Date

2020-02-24

DOI

10.1186/s12859-020-3368-2

Peer reviewed

SOFTWARE

Open Access

GenEpi: gene-based epistasis discovery using machine learning



Yu-Chuan Chang^{1,2}, June-Tai Wu³, Ming-Yi Hong⁴, Yi-An Tung^{2,5}, Ping-Han Hsieh¹, Sook Wah Yee⁶, Kathleen M. Giacomini^{6,7}, Yen-Jen Oyang¹, Chien-Yu Chen^{2,4*}  and for the Alzheimer's Disease Neuroimaging Initiative

Abstract

Background: Genome-wide association studies (GWAS) provide a powerful means to identify associations between genetic variants and phenotypes. However, GWAS techniques for detecting epistasis, the interactions between genetic variants associated with phenotypes, are still limited. We believe that developing an efficient and effective GWAS method to detect epistasis will be a key for discovering sophisticated pathogenesis, which is especially important for complex diseases such as Alzheimer's disease (AD).

Results: In this regard, this study presents GenEpi, a computational package to uncover epistasis associated with phenotypes by the proposed machine learning approach. GenEpi identifies both within-gene and cross-gene epistasis through a two-stage modeling workflow. In both stages, GenEpi adopts two-element combinatorial encoding when producing features and constructs the prediction models by L1-regularized regression with stability selection. The simulated data showed that GenEpi outperforms other widely-used methods on detecting the ground-truth epistasis. As real data is concerned, this study uses AD as an example to reveal the capability of GenEpi in finding disease-related variants and variant interactions that show both biological meanings and predictive power.

Conclusions: The results on simulation data and AD demonstrated that GenEpi has the ability to detect the epistasis associated with phenotypes effectively and efficiently. The released package can be generalized to largely facilitate the studies of many complex diseases in the near future.

Keywords: GWAS, Epistasis, Machine learning

Background

Genome-wide association studies (GWAS) is a univariate examination of a genome-wide set of genetic variants to determine if any single variant is associated with the phenotype of interest [1]. The first GWAS was published in 2002 [2], and 3 years later, the most remarkable GWAS regarding age-related macular degeneration

(AMD) was published [3]. Their study investigated the association of 105,980 single nucleotide polymorphisms (SNPs) with AMD on 96 cases and 50 control subjects. This study showed that the SNPs in the complement factor H (CFH) gene, including a non-synonymous SNP, are significantly associated with AMD. Up to 2019, there have been more than hundreds of thousands individuals being studied in typical GWAS protocols, and over 210,498 variant-disease associations between 117,337 SNPs and 10,358 phenotypes have been discovered [4]. These studies demonstrated the potential of GWAS to identify genetic variants associated with many categories of phenotypes, including risks for diseases such as various cancers, and variations in therapeutic and adverse responses to drugs. However, the success of univariate GWAS is limited to monogenic phenotypes (e.g. Mendelian

* Correspondence: chienyuchen@ntu.edu.tw

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A comprehensive list of consortium members appears at the end of the paper.

²Taiwan AI Labs, Taipei, 10351 Taiwan

⁴Department of Biomechanics Engineering, National Taiwan University, Taipei, 10617 Taiwan

Full list of author information is available at the end of the article



diseases). The impact of variant interactions, also known as epistasis on the formation of diseases [5] is often underestimated in traditional GWAS analysis [6–8].

A major limitation of traditional GWAS is that it considers only one genetic variant at a time, and ignores underlying epistasis of variants that might have stronger associations [9]. Researchers have found that GWAS has limitation in identifying the association in complex diseases [10, 11]. Easton et al. suggested that a number of susceptible loci identified by GWAS usually have very small effect sizes [12]. Studies have also demonstrated that the existence of epistasis is an important factor contributing to phenotypes, especially in complex diseases such as hypertension, diabetes and obesity [11]. Therefore, developing analytical methods to identify epistasis efficiently is critical to understanding the genetic factors [8, 13], and has attracted a wide range of research interests in recent years [7, 14].

There are, however, two main challenges to discover epistasis: computational complexity and statistical power [15]. The first challenge results from the curse of dimensionality. When more genetic variants are considered, the number of interactions increases exponentially. Based on the specification of a major commercial technology, Illumina Arrays, a whole-genome array can investigate over 4 million markers per sample simultaneously. In order to evaluate the pairwise interactions from this microarray, about 8×10^{12} statistical tests need to be processed. Even though Marchini et al. have demonstrated that pairwise interactions of 3×10^5 loci is computationally possible with currently available computational resources, it still remains challenging when the Illumina Arrays are considered [16]. The second challenge is the issue of statistical power. Since a huge number of statistical tests are conducted on a limited sample size with high-dimensional interactions, many false positives arise by random chances. In recent years, new methods have been developed to tackle the issue of epistasis [11, 17]. Statistical approaches include FastEpistasis [18] and BOOST [19]; both of them has been included in a well-known GWAS software called PLINK [20, 21]. Machine learning approaches such as Multifactor Dimensionality Reduction [22], ReliefF [23], random forest-like algorithms [24–26] and other methodologies have also been developed for detecting epistasis [17].

Since the biological experiments used to validate these methodologies are still in demand, there are no standard analysis methods for epistasis despite the rapid improvement in computational performance. In 2016, Murk used FastEpistasis and BOOST to search SNP-SNP interactions on a huge dataset called Genetic Epidemiology Research on Adult Health and Aging (GERA) that included 78,486 subjects, but still failed to detect a significant and replicable interaction after exhaustively searching through 45 billion possible interactions for 10 complex diseases of interest

[27]. Alzheimer's disease (AD) is one of the most important complex diseases and its pathogenesis, which clearly has a genetic basis, is still ill-defined. In 2014, Sage Bionetworks held a competition called The Dialogue for Reverse Engineering Assessments and Methods Challenge (DREAM Challenge) for AD, which tried to use crowdsourcing to assess the capability of current computational methods to predict the change in cognitive examination based on genetic data. However, no significant contribution of genetic features except the *APOE* haplotype to the predictive performance was observed by any competition teams [28]. In order to discover more SNP interactions with both statistical and biological significance, this study presents GenEpi, a package to reveal epistasis related to the phenotype using machine learning and introduces the application of GenEpi on AD.

Implementation

The architecture of GenEpi is shown in Fig. 1. GenEpi is designed to group SNPs by a set of loci in the genome. For examples, a locus could be a gene. In other words, we use gene boundaries to group SNPs. A locus can be generalized to any particular regions in the genome, e.g. promoters, enhancers, etc. GenEpi first considers the genetic variants within a particular region as features in the first stage, because it is believed that SNPs within a functional region might have a higher chance to interact with each other and to influence molecular functions. The idea of within-gene epistasis analysis followed by cross-gene analysis is not new, which has also been used in previous studies [29–32]. Differently, GenEpi adopts two-element combinatorial encoding when producing features and models them by L1-regularized regression with stability selection, which will be explained in Section 2.3. In the first stage (STAGE 1) of GenEpi, the genotype features from each single gene will be combinatorially encoded and modeled independently by L1-regularized regression with stability selection. In this way, we can estimate the prediction performance of each gene and detect within-gene epistasis with a low false positive rate. In the second stage (STAGE 2), both of the individual SNP and the within-gene epistasis features selected by STAGE 1 are pooled together to generate cross-gene epistasis features, and modeled again by L1-regularized regression with stability selection as STAGE 1. Finally, the user can combine the selected genetic features with environmental factors such as clinical features to build the final prediction models. In addition to the main procedures, two pre-processing steps are also implemented in GenEpi: retrieving the gene information from public databases and reducing the gene information from public databases and reducing the dimensionality of the features using linkage disequilibrium (LD). In the end, we released a Python package that implements GenEpi. The details of these steps and the GenEpi method will be described in the following sections.

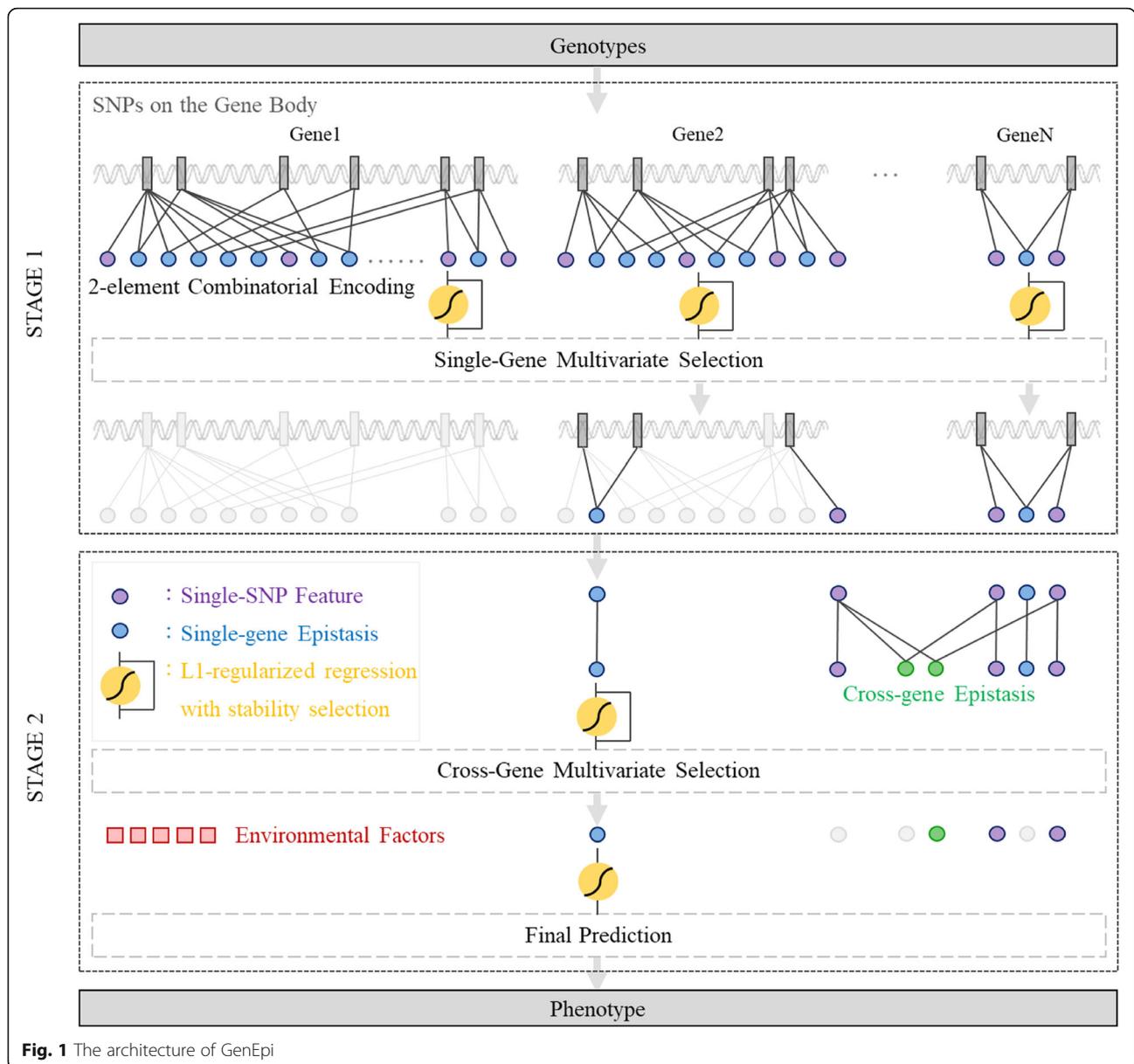


Fig. 1 The architecture of GenEpi

University of California Santa Cruz (UCSC) database

To obtain the gene information such as official gene symbols and genomic coordinates, we retrieved kgXref and knownGene data table from the UCSC human genome annotation database [33, 34]. The version of the database we used is the Feb. 2009 assembly of the human genome hg19, GRCh37 Genome Reference Consortium Human Reference 37. The two data tables were merged in order to generate a local database containing the gene symbols as well as the genomic coordinates of each gene. The in-house script we built could update this local database automatically. It is noted that there are many different categories of genes in the RefSeq database. In this study, we only focused on the mRNA and non-coding RNA (22,376 genes in total). The

selected transcripts were projected on the genomic coordinates and the coordinates of corresponding genes were determined based on the leftmost and rightmost positions of the corresponding transcripts. Moreover, to discover the factors that might affect the transcription of genes, we also retained the promoter region of each gene. In genetics, the promoter region is a segment of DNA that initiates the transcription of a particular gene. Promoters are located near the transcription start sites of genes, on the upstream of the same DNA strand (towards the 5' region of the sense strand of the transcript). In general, a promoter region can be 100–1000 base pairs long. In this study, we extracted 1000 nucleotides on the upstream of the starting position of each gene as the promoter region.

Estimation of linkage disequilibrium

In GWAS datasets, a SNP often exhibits high dependency with its nearby SNPs because of linkage disequilibrium (LD). In the practical implantation, we prefer to group these dependent features to reduce the dimension of features. In other words, we can take the advantages of LD to reduce the dimensionality of SNP features. In this regard, we adopted the same approach developed by Lewontin [35] to estimate LD (see Additional file 1 Section S.1). We used $D' > 0.9$ and $r^2 > 0.9$ as the criteria to group highly dependent SNP features as blocks. In each block, we chose the features with the largest minor allele frequency to represent other features in the same block. It is important to look at the SNPs falling in the same LD blocks with the SNPs discovered by GenEpi. Some true interactions might be skipped owing to some strong signals provided by the SNPs in the same LD block.

Discovery of within-gene epistasis

The main objective of the first stage in GenEpi is to select candidate features from each gene. In order to extract SNP features for a gene, we used the start and end positions of each gene from the local UCSC database to split the SNP features after dimension reduction. Since there are 22,376 genes in the UCSC database, we obtained 22,376 subsets of the SNP features. In each subset, a SNP feature with the alleles 'A' and 'a' could have three possible genotypes, AA, Aa and aa, which are used to refer to the pairs of alleles. The pairs of alleles are subsequently separated into three binary features using one-hot encoding. In order to evaluate epistasis, we generated interacting features by crossing each pair of genotype features. Considering the false positive rate and computational complexity, we only focused on pairwise interactions of epistasis throughout this study. We defined the interaction between two SNPs in Eq. 1. In Eq. 1, $\alpha_1 SNP_1 + \alpha_2 SNP_2$ stand for the additive interactions and $\alpha_{int(1,2)} SNP_1 \otimes SNP_2$ represents the synergistic interactions that contain nine terms.

$$\begin{aligned}
 y &= \alpha_0 + \sum_{m \in \{AA, Aa, aa\}} \alpha_{1,m} SNP_{1,m} \\
 &+ \sum_{m \in \{AA, Aa, aa\}} \alpha_{2,m} SNP_{2,m} \\
 &+ \sum_{m,n \in \{AA, Aa, aa\}} \alpha_{1,m,2,n} SNP_{1,m} SNP_{2,n} \\
 &= \alpha_0 + \alpha_1 SNP_1 + \alpha_2 SNP_2 + \alpha_{int(1,2)} SNP_1 \otimes SNP_2
 \end{aligned} \tag{1}$$

Before modeling each subset of genotype features, two criteria were adopted to exclude low quality data. The

first criterion is that the genotype frequency of a feature should exceed 5%, where the genotype frequency means the proportion of a genotype among the total samples in the dataset. The second criterion is regarding the association between the feature and the phenotype. We used χ^2 test to estimate the association between the feature and the phenotype, and the p -value should be smaller than 0.01. In the end, a gene may have multiple SNPs. The general form of the linear model for a gene with k SNPs is defined as Eq. 2, which is termed as two-element combinatorial encoding.

$$\begin{aligned}
 y &= \alpha_0 + \sum_{i=1}^k \alpha_i SNP_i \\
 &+ \sum_{i=1}^k \sum_{j=1 \wedge j \neq i}^k \alpha_{int(i,j)} SNP_i \otimes SNP_j
 \end{aligned} \tag{2}$$

We conducted L1-regularized regression [36] with stability selection [37] for modeling each gene. The sparsity of the L1-regularized model prefers solutions with a smaller number of features, which effectively reduces the number of features. As in Equation 3, L1-regularized regression uses an additional regularization term $\lambda \|\alpha\|_1$ to restrict the weight of each feature by shrinking some of them to 0 so that the non-zero remainders can represent the exact set of true features when given a proper λ . In Equation 3, we have the vector $\mathbf{SNP} = (SNP_1, \dots, SNP_i, \dots, SNP_k, SNP_1 \otimes SNP_2, \dots, SNP_i \otimes SNP_j, \dots, SNP_{k-1} \otimes SNP_k)$, the corresponding coefficients $\alpha = (\alpha_1, \dots, \alpha_i, \dots, \alpha_k, \alpha_{int(1,2)}, \dots, \alpha_{int(i,j)}, \dots, \alpha_{int(k-1, k)})$, the target y_l takes the values $\{-1, 1\}$ at sample l and c is a constant to be determined during modeling.

$$\alpha^\lambda = \min_{\alpha, c} \sum_{l=1}^n \log(\exp(-y_l \times (\mathbf{SNP}_l^T \alpha + c)) + 1) + \lambda \|\alpha\|_1 \tag{3}$$

, where n stands for the number of samples. It should be noticed that, if the features are conditional dependent, the solution of these equations will not be unique. It would lose generality to determine the proper amount of λ when we only consider a possible solution of weight vector α . Resampling is an intuitive technique to increase the generality, which can largely reduce the false positive rate. Here, we used stability selection [37] to tackle this problem. Stability selection works by resampling and remodeling the training set hundreds of times, followed by picking out the features that are repeatedly selected across randomization. In this study, we executed this randomization 500 times, and the features selected by stability selection would be retained for the next stage.

Discovery of cross-gene epistasis

In the second stage, we used the features selected by STAGE 1 to generate cross-gene epistasis features. To avoid missing any possible association between genotype features and phenotype. In the default setting of the GenEpi package, we include all the genes with non-zero F1 score to go into the next stage. Then we applied the same selection procedure described in Section 2.3 to find the cross-gene epistasis that are associated with the phenotype. The procedures were slightly modified here. Since we only focused on pairwise interactions, instead of using the entire features we selected in STAGE 1, we only used single-SNP features to generate cross-gene epistasis features. Also, we used the genotype frequency and the p -values of χ^2 test to control the quality of features and to avoid overfitting. Nevertheless, the p -value of each feature in this stage should be smaller than 10^{-5} . All of the features from different genes would be merged for modeling cross-gene epistasis. We conducted L1-regularized regression for modeling, and the stability selection were used once again to select the final genotype feature set. Since the phenotype may also be affected by environmental factors, after determining the final set of genotype features, the user can included the environmental factors such as clinical assessments for constructing the final model. Subsequently, the final model was evaluated through a process called double cross validation (CV). In the external loop of double CV, all the instances were divided into two subsets to serve as training and independent test sets. In this study, we used 2-fold CV and leave-one-out CV (LOO CV) in external loop for evaluation. In the internal loop, we also used 2-fold CV for model selection.

Materials

This study applied GenEpi on an AD cohort, which was used in Alzheimer's disease Dream Challenge [28]. In total, the cohort consists of 767 participants, who were healthy elderly, mild cognitive impairment (MCI) and AD patients from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The 767 ADNI participants consist of 241 cognitively normal (CN), 130 Early MCI (EMCI), 273 Late MCI (LMCI) and 123 AD participants. According to the definition of the four categories used in the ADNI database, the samples of AD are in same stage. We adopted only genetic features in this study. All the genetic data has been pre-processed by the organizers that held the challenge [28]. The genetic data were genotyped using the Illumina Human610-Quad BeadChip and Illumina HumanOmniExpress BeadChip. The multidimensional scaling analysis was applied by PLINK using HAPMAP3 to ensure that samples are within the cluster of European populations. Subsequently, the data were imputed according to the 1000

genome haplotypes. After imputation, there were 12,809, 667 genotype features in total. For predicting the diagnosis of AD, we used 364 participants, of which the clinical diagnosis are CN or AD, to predict which samples are control subjects or the AD patients.

Results

This study compared GenEpi with several commonly used algorithms for detecting epistasis, including FastEpistasis, BOOST and ReliefF. The simulation data demonstrated that GenEpi outperforms the other methods in ranking the true epistasis as the top one. As real data is concerned, the results suggest that the epistasis selected by GenEpi has the best predictive power for diagnosis of AD. The proposed model of predicting AD contains 14 genetic features, including 24 SNPs from 12 genes that contain the well-known causal gene, APOE. The 2-fold cross validation (CV) and leave-one-out CV (LOO CV) accuracy of this model are 0.829 and 0.832, respectively. The results on AD demonstrated that GenEpi has the ability to detect the epistasis associated with the phenotype effectively and efficiently. The released package can be generalized to largely facilitate the studies of many complex diseases in the near future.

We will demonstrate our experiments in following three parts of this section. In the first part, we applied GenEpi and other algorithms for detecting epistasis, including FastEpistasis [18], BOOST [19] and ReliefF [23, 38] on simulation data for validation and comparison. In the second part, we applied GenEpi on the ADNI dataset to categorize each sample as control subjects or AD patients, evaluated by precision, recall, accuracy and F1 score ($2 \times (\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$). In the final part, we compared GenEpi with other algorithms on the ADNI dataset in terms of computing time and prediction performance on real data.

Experiments on simulation data

We applied different algorithms on simulation data for validation and comparison. All of the simulation datasets are generated by the simulator GAMETES [39], which is publicly accessible on the web site <https://popmodels.cancercontrol.cancer.gov/gsr/packages/gametes/>. We designed two types of simulation datasets: basic and complex models. The 'Model 1', 'Model 2' and 'Model 3' are simulation datasets with the basic model, which means that each dataset contained only one epistasis consisting of a SNP pair. All of the basic-model datasets are in the same setting as follows: #individuals = 2000, case/control ratio = 1, #SNPs = 100, #replicates = 100, minor allele frequency of target SNPs = 0.2, and heritability = 0.2. The complex model means one dataset contains multiple epistasis from different SNP pairs. Here, the 'Combined

Model 1+2+3' is a complex model dataset containing three epistasis from the previous three basic models.

Figure 2 provided the results of these four simulation datasets. Figure 2a shows that the ranking of the target epistasis reported by GenEpi in the 100 replicates of each basic-model dataset are always ranked as the first. In contrast, for FastEpistasis and BOOST, the medians of the ranking of the target epistasis among the 100 runs of simulation are one but the averages are not. The number of failures of FastEpistasis and BOOST in 100 replicates of three basic models are 6, 1, 16 and 5, 1, 14, respectively. For the result of the complex model dataset in Fig. 2b, the superiority of GenEpi over other algorithms is more obvious. In the 100 runs of simulation, GenEpi reported the three target epistasis as the top three important features every time. In contrast, FastEpistasis and Boost failed to report the three target epistasis as the top three important features consistently.

When ReliefF was compared, since the Python package scikit-rebate [38] that we used for implementing ReliefF only reports the importance of individual SNPs instead of the scores for epistasis (SNP pairs), we listed the medians of ReliefF's ranking for each SNP in the target epistasis in Table 1. Table 1 reveals that ReliefF can detect the SNPs in the target epistasis in the basic models,

but failed to report the three target epistasis as the top three important features in the complex model dataset.

The superiority of GenEpi is owing to the proposed two-element combinatorial encoding of the genotype features and the L1-regularized regression with stability selection. In contrast with other statistical algorithms such as FastEpistasis and BOOST, which only evaluate the epistasis of a SNP pair one at a time, GenEpi considers interactions between combinatorial features by multivariate models. Moreover, the false positives among the epistasis can be filtered out by resampling and remodeling the dataset hundreds of times. To evaluate the effect of stability selection, we applied both L1-regularized regression with and without stability selection on the complex model dataset to compare the number of false positives, which is defined as the number of non-target epistasis in the final output of GenEpi. As shown in Fig. 3, stability selection can reduce the mean false positive rate effectively and minimize the variance of false positive rate as well.

Classifying AD patients

In predicting control subjects or AD patients, we applied GenEpi on the 364 samples with CN (as control) or AD. After dimensionality reduction, 12,102,888 out of the 12,

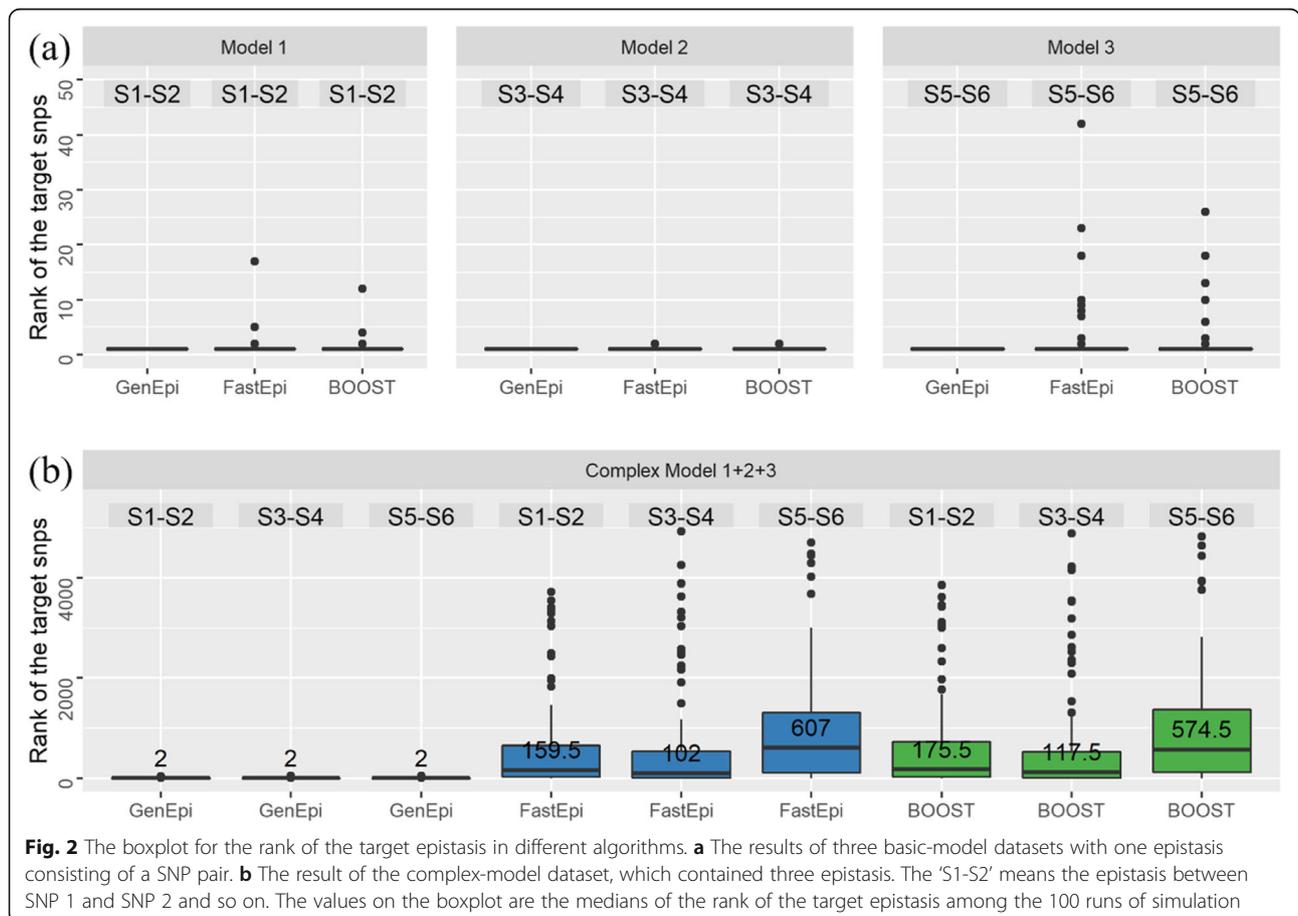


Table 1 The medians of the rank of the SNPs in the target epistasis for ReliefF

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
Basic Model	1	2	1	2	1	2
Complex Model	7	8.5	9.5	11.5	11	19.5

809,667 SNPs in the ADNI dataset were retained, and 4,916,249 of them are located in 20,206 genes (Additional file 1 Table S1). In the step 4 of selecting epistasis, there are 34,689 genetic features selected and 765 of them are single SNP features, while the other 33,924 are epistasis features within genes. The final model contained 14 genetic features, including 24 SNPs from 12 genes. These features contained two single SNP features, 11 within-gene epistasis features and one cross-gene epistasis feature. As shown in Table 2, the 2-fold cross validation (CV) and leave-one-out CV (LOO CV) accuracy of this model are 0.83 and 0.83, respectively.

We listed the statistical significance of the selected genetic features in Table 3. The first column lists each feature by its RSID (Reference SNP cluster ID) and the genotype (denoted as RSID_genotype), the pairwise epistasis features are represented using two SNPs. The last column describes the genes where the SNPs are located according to the genomic coordinates. We used a star sign to denote the epistasis between genes. Here, only the feature (rs3130614_BB, rs41276317_AB) is cross-gene epistasis (for MICB and TOB2). The weights in the second column were extracted from the linear model we defined in Section 2.4. The signs of the weights indicate if a feature is a causal or protective genotype, which is consistent with the corresponding odds ratio. The *p*-value of the χ^2 test showed that these features are significantly associated with the phenotype.

Comparison with different algorithms

In this section, we compared GenEpi with other algorithms for detecting epistasis, including FastEpistasis [18], BOOST [19] and ReliefF [23] in terms of computing time and prediction performance. We used Microsoft Azure E32 v3 as the computing resource, which

Table 2 The score of different models in predicting control subjects or AD patients

	Precision	Recall	F1 Score	Accuracy
Training	0.9633	0.8537	0.9052	0.9396
2-fold CV	0.7748	0.6992	0.7350	0.8297
LOO CV	0.8100	0.6585	0.7265	0.8324

F1 Score = $2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$; 'Training' stands for the process of a single-loop CV; '2-fold CV' means that 2-fold CV was used in the external loop of double CV; 'LOO CV' means that LOO CV was used in the external loop of double CV

contains 32 CPUs and 256 GB RAM. Since the PLINK (version 2.0) has imported FastEpistasis and BOOST, we used PLINK to test these two algorithms. For ReliefF, we employed a Python package called scikit-rebate [38] for implementation. Among these algorithms, only FastEpistasis can afford the computation of the whole set of SNPs. In this regard, 12,809,667 SNPs were used by FastEpistasis (Table 4). On the other hand, GenEpi only focuses on the SNPs in the gene regions. In this regard, the number of input SNPs for estimating epistasis reduced to 4,916,249. BOOST took the same subsets of SNPs as GenEpi (Table 4). When taking the same subset of SNPs as GenEpi and BOOST, ReliefF still caused memory errors. Therefore, we used the subsets of SNPs that selected by STAGE 1 of GenEpi as the input of ReliefF, which are 33,868. We selected the top 15, 30, 45 and 60 rankings from the results of these algorithms for comparing the prediction performance, and used L1-regularized regression to build the models for classifying AD patients for comparison. Table 4 shows that GenEpi is an efficient method, which can deliver satisfied results for the epistasis discovery of 4 millions of SNPs within 9.95 CPU-days. The comparison of execution time is unfair to FastEpistasis, since FastEpistasis used the whole set of SNP, which is about 2.6 time larger than the subset of it be used in GenEpi. When accuracy is considered, GenEpi has the best prediction performance despite the fact that GenEpi only uses the subset of SNPs from the final model. GenEpi shows that the time needed for identifying epistasis can be drastically reduced, without compromise to the performance. We

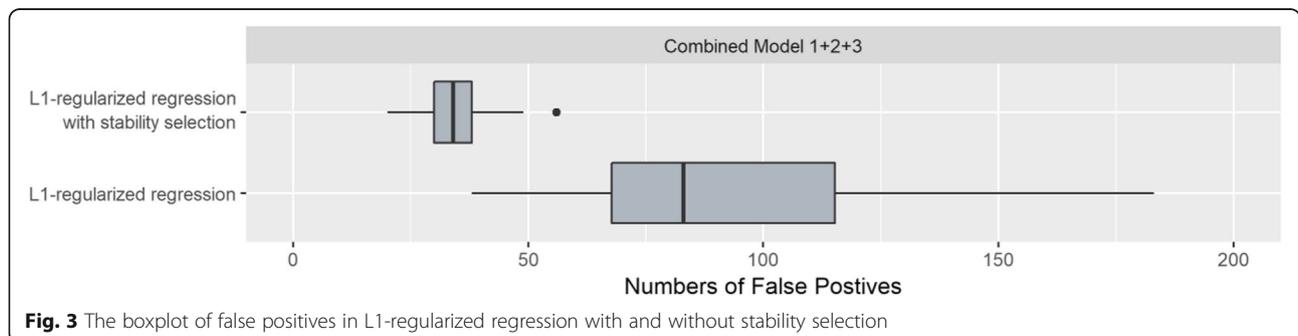


Fig. 3 The boxplot of false positives in L1-regularized regression with and without stability selection

Table 3 The statistical significance of genetic features selected by GenEpi in predicting patients with AD

Selected SNPs (RSID)	Weight	Odds Ratio	χ^2 -test P-value	Genotype Frequency	Gene
rs3130614_BB, rs41276317_AB	3.16	19.23	1.42E-09	0.0742	<i>MICB</i> ^{ab} <i>TOB2</i>
rs12095538_BB, rs2774308_AB	2.41	7.69	6.87E-07	0.0824	<i>SYT6</i>
rs12926153_AB, rs12922908_AA	1.18	4.83	6.89E-07	0.1511	<i>CLEC16A</i>
rs9652600_AB, rs12922908_AA	0.94	4.83	6.89E-07	0.1511	<i>CLEC16A</i>
rs9344977_BB, rs56148686_AB	1.94	4.32	1.14E-06	0.1813	<i>BACH2</i>
rs429358_AA	-2.01	0.17	1.73E-06	0.5962	<i>APOE</i>
rs56233035_AB, rs3678_AB	2.26	10.16	1.91E-06	0.0604	<i>CACNA1E</i>
rs11675339_AA, rs2710687_AA	2.32	3.94	3.55E-06	0.1923	<i>VSNL1</i>
rs12189429_BB, rs6881360_AA	1.36	4.34	3.65E-06	0.467	<i>ADAMTS12</i>
rs12187423_BB, rs6881360_AA	0.58	4.34	3.65E-06	0.467	<i>ADAMTS12</i>
rs10831829_BB, rs12366151_AA	3.48	9.50	4.90E-06	0.0577	<i>PARVA</i>
rs2052573_BB, rs34580133_AB	1.80	4.08	5.00E-06	0.1648	<i>LINC00299</i>
rs2421701_AB, rs200512701_AB	1.82	4.12	5.29E-06	0.1593	<i>TNKS2</i>
rs769449_AA	-1.19	0.16	8.42E-06	0.6648	<i>APOE</i>

The sign ^{ab} between two gene symbols indicates cross-gene epistasis

provided the ROC curves for the classification task in Fig. 4, and it shows that GenEpi achieved the best performance in double 2-fold CV procedures, of which the area under the curve (AUC) is 0.85.

Discussion

The results in the previous section revealed the power of GenEpi to identify phenotype-associated epistasis efficiently. GenEpi selected 14 features from 12 genes to categorize patients with AD. Since AD is a chronic neurodegenerative disease, our findings would be supported if the gene identified by GenEpi are expressed in brains. We downloaded the median RPKM by tissue dataset (GTEx Analysis V6: dbGaP Accession phs000424.v6.p1) of the GTEx Project [40] and plotted a heatmap to inspect the gene expression of these genes in different tissues, as shown in Fig. 5. Among the 12 genes selected by GenEpi, 11 have high expression level in the brain tissues. In addition, five genes, *CLEC16A*, *VSNL1*, *SYT6*, *CACNA1E* and *LINC00299*, have a similar expression pattern with *APOE*. The R script for drawing the heatmap for GTEx dataset could be found in Additional file 2.

Table 4 The comparison of different algorithms

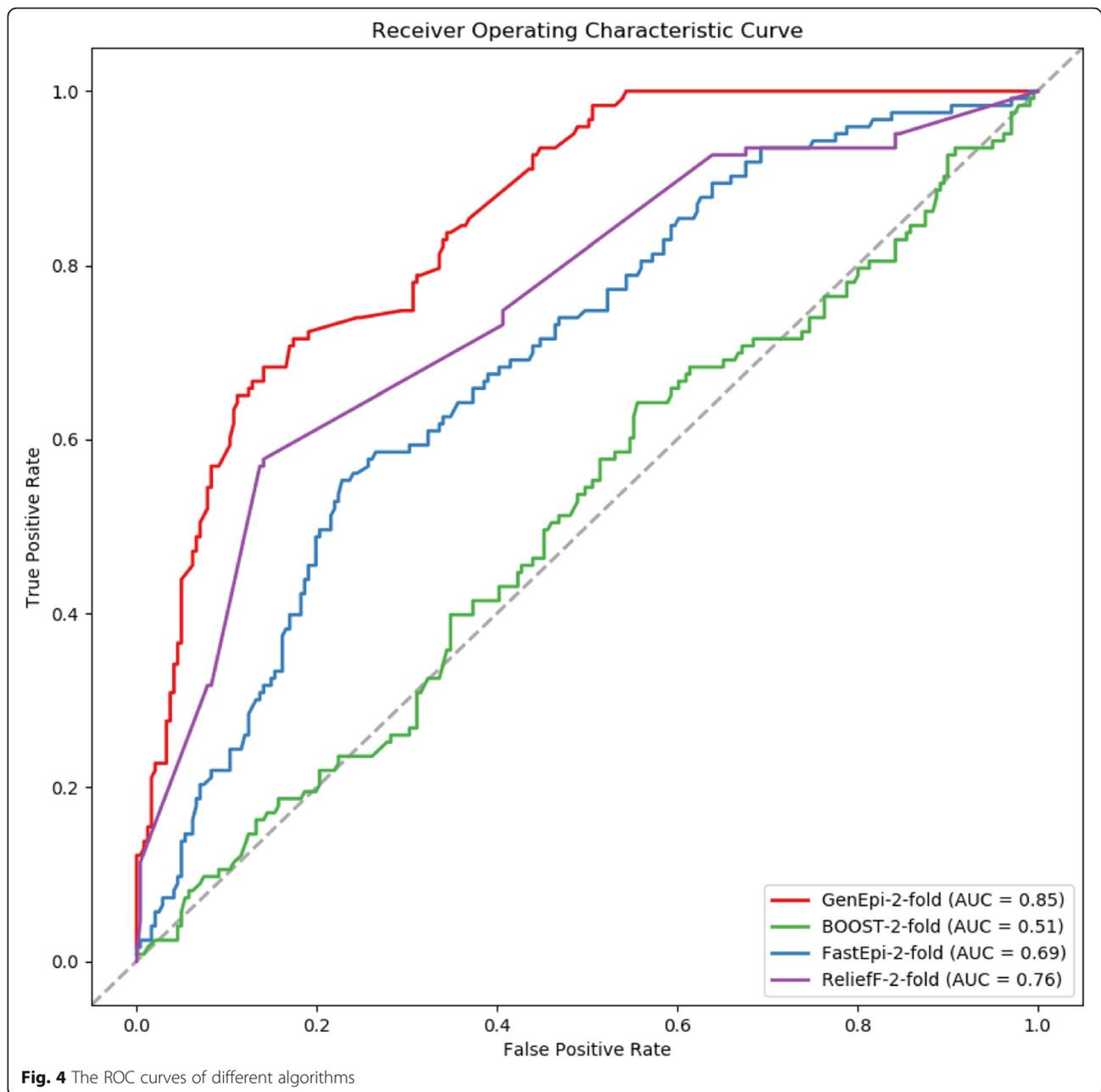
Algorithm	# Input SNP	Time Cost	Top 15	Top 30	Top 45	Top 60
GenEpi	4,916,249	9.95	0.76	0.72	0.71	0.68
BOOST	4,916,249	2157.6	0.31	0.24	0.30	0.37
ReliefF	33,868	0.11	0.52	0.48	0.45	0.46
FastEpistasis	12,809,667	836.8	0.62	0.61	0.60	0.59

'Time Cost' is the time spent on identifying the epistasis, which was measured by single CPU time in days. The values in column top 15, top 30, top 45 and top 60 are the 2-fold CV scores. The 2-fold CV scores are the F1 scores

These 12 genes are categorized as cross-gene epistasis, single-gene epistasis and single-SNP features based on the feature types selected by GenEpi. GenEpi detected only one cross-gene epistasis, which is *MICB* * *TOB2*. We found several evidences to demonstrate that this interaction might have true association with AD (see Additional file 1 Section S.2.1).

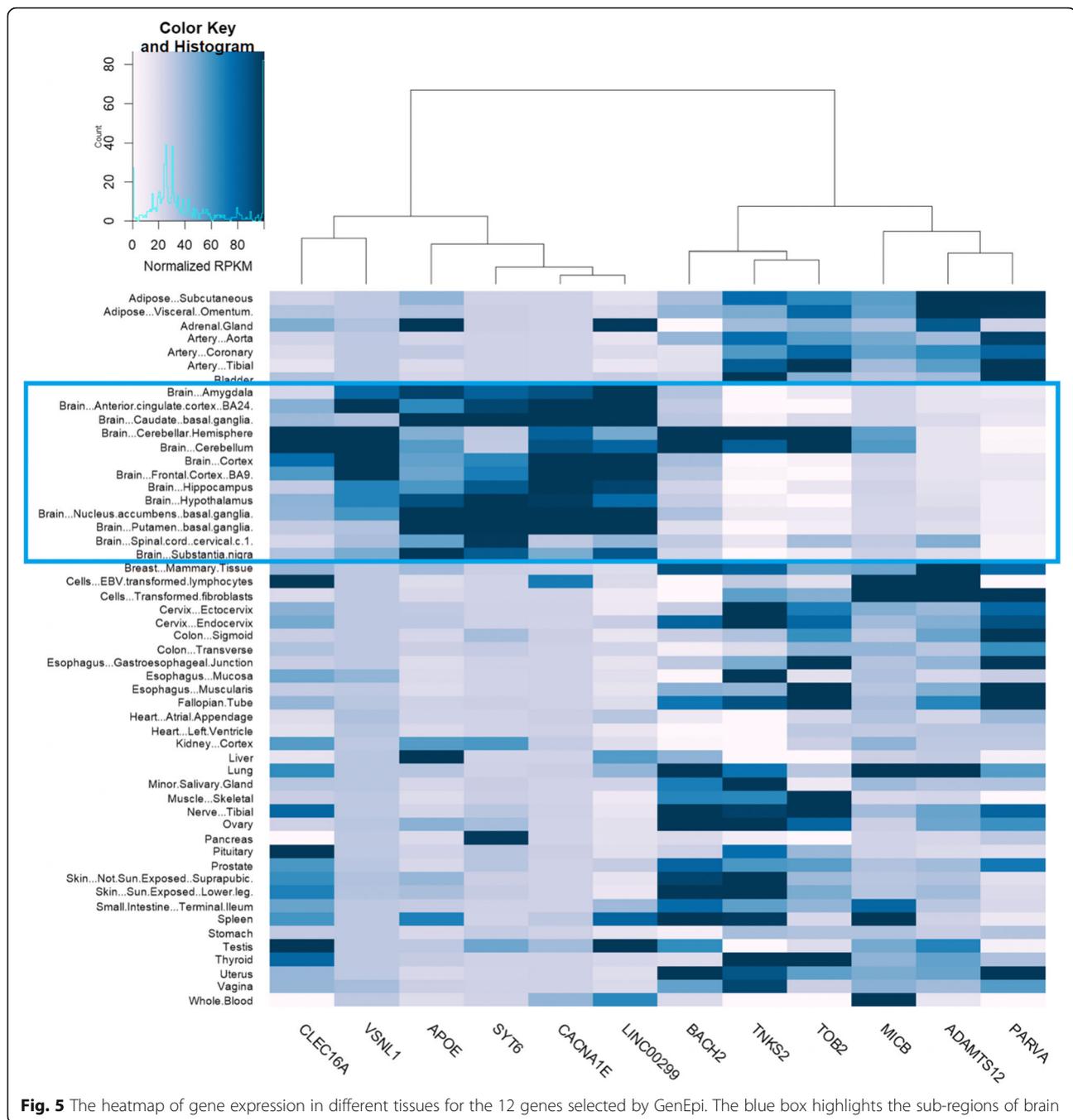
About the 11 single-gene epistasis, there are several possible reasons accounting for intramolecular SNP-SNP interactions identified in this study. The first is a synergistic regulation of transcription [41], the second is a synergistic interaction between transcriptional and post-transcriptional regulation [42], and the third is an intramolecular SNP pair modulating the expression of two separate neighboring genes [43]. Most of the single-gene epistasis selected by GenEpi can be explained by these three possible reasons (see Additional file 1 Section S.2.2) and only two of the SNP-SNP interactions are not immediately clear at this moment. Last, there are only two single-SNP features and both of them are located in *APOE*, which is a well-known causal gene of AD, revealing that GenEpi is an effective tool to identify disease-causing genes. Moreover, GenEpi successfully selected out the SNP rs429358, which determines the allele type of *APOE* with rs7412.

While GenEpi has shown its ability to identify epistasis efficiently, it might still has the following limitations. Firstly, GenEpi can only detect pairwise interactions. Considering the false positive rate and computational complexity, it may not be appropriate for continuously generating the high-dimension interactions. A feature engineering-free method such as deep learning could be applied for discovering the high-dimension interactions.



Second, GenEpi is a memory-consuming package, which might cause memory errors when calculating the epistasis of a gene containing a large number of SNPs. We recommend that the memory for running GenEpi should be over 256 GB. Since most of features may not be associated with the phenotype, additional filters for feature selection can be designed to further reduce the number of features before modeling. Finally, a small sample size may lead overfitting, which forces us to use strict thresholds during feature selection. In this way, GenEpi delivers a high precision rate, but might suffer having false

negatives. This implies different GWAS data might detect different sets of true positives. In traditional GWAS, meta-analysis [44] can be used to identify the common effects from multiple studies. This post statistical procedure could be considered for obtaining a common set from multiple GWAS data. In summary, the results of this study demonstrated that GenEpi is a promising software package to identify causal SNPs and epistasis in GWAS, and it can be further used to predict the phenotypes. With the demonstrated efficiency, GenEpi is a powerful tool to explore gene-gene interactions that underlie complex diseases.



Conclusions

This study presents GenEpi, a computational package to uncover epistasis associated with phenotypes by the proposed machine learning approach, which adopts two-element combinatorial encoding when producing features and constructs the prediction models by L1-regularized regression with stability selection. The results on simulation data and AD demonstrated that GenEpi has the ability to

detect the epistasis associated with phenotypes effectively and efficiently. Furthermore, the release package GenEpi is an open-source Python package and available free of charge for non-commercial users. The package has been published on The Python Package Index, and GitHub (<https://github.com/Chester75321/GenEpi>), can be generalized to largely facilitate the studies of many complex diseases in the near future.

Availability and requirements

Project name: GenEpi.

Project home page: <https://github.com/Chester75321/GenEpi>

Operating system(s): Platform independent.

Programming language: Python.

License: MIT license.

Any restrictions to use by non-academics: license needed.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12859-020-3368-2>.

Additional file 1. Supplementary information for literature survey of the genetic features selected by GenEpi and the formulas for linkage disequilibrium estimation.

Additional file 2. The R script to draw a heatmap for GTEx dataset.

Abbreviations

AD: Alzheimer's Disease; ADNI: Alzheimer's Disease Neuroimaging Initiative; AMD: Age-related Macular Degeneration; AUC: Area Under the Curve; CFH: Complement Factor H; CN: Cognitively Normal; CV: Cross Validation; DREAM Challenge: The Dialogue for Reverse Engineering Assessments and Methods Challenge; EMCI: Early Mild Cognitive Impairment; GERA: Genetic Epidemiology Research on Adult Health and Aging; GWAS: Genome-Wide Association Studies; LD: Linkage Disequilibrium; LMCI: Late Mild Cognitive Impairment; LOO CV: Leave-One-Out Cross Validation; MCI: Mild Cognitive Impairment; RSID: Reference SNP cluster ID; SNP: Single Nucleotide Polymorphisms; UCSC: University of California Santa Cruz

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. This research is supported by the National Natural Science Foundation of China (Nos 61422204, 61473149, 61501230), the NUAU Fundamental Research Funds (No. NE2013105), the Jiangsu Qinglan Project of China. At Indiana University, this work was supported by NIH R01 LM011360, U01 AG024904, RC2 AG036535, R01 AG19771, P30 AG10133, UL1 TR001108, R01 AG 042437, and R01 AG046171; NSF IIS-1117335; DOD W81XWH-14-2-0151, W81XWH-13-1-0259, and W81XWH-12-2-0012; NCAA 14132004; and CTSI SPARC Program.

Consortia

Alzheimer's Disease Neuroimaging Initiative.

Michael W. Weiner⁷, Paul Aisen⁶, Ronald Petersen^{9,10}, Clifford R. Jack Jr.¹⁰, Sara S. Mason¹⁰, Colleen S. Albers¹⁰, David Knopman¹⁰, Kris Johnson¹⁰, William Jagust¹¹, John Q. Trojanowski¹², Arthur W. Toga¹³, Laurel Beckett¹⁴, Robert C. Green¹⁵, Martin R. Farlow¹⁵, Ann Marie Hake¹⁵, Brandy R. Matthews¹⁵, Jared R. Brosch¹⁵, Scott Herring¹⁵, Cynthia Hunt¹⁵, Leslie M. Shaw¹⁶, Beau Ances¹⁶, John C. Morris¹⁶, Maria Carroll¹⁶, Mary L. Creech¹⁶, Erin Franklin¹⁶, Mark A. Mintun¹⁶, Stacy Schneider¹⁶, Angela Oliver¹⁶, Jeffrey Kaye¹⁷, Joseph Quinn¹⁷, Lisa Silbert¹⁷, Betty Lind¹⁷, Raina Carter¹⁷, Sara Dolen¹⁷, Lon S. Schneider¹³, Sonia Pawluczky¹³, Mauricio Beccera¹³, Liberty Teodoro¹³, Bryan M. Spann¹³, James Brewer¹⁸, Helen Vanderswag¹⁸, Adam Fleisher^{18,19}, Pierre Tariot¹⁹, Anna Burke¹⁹, Nadira Trncic¹⁹, Stephanie Reeder¹⁹, Judith L. Heidebrink²⁰, Joanne L. Lord²⁰, Rachelle S. Doody²¹, Javier Villanueva-Meyer²¹, Munir Chowdhury²¹, Susan Rountree²¹, Mimi Dang²¹, Yaa-kov Stern²², Lawrence S. Honig²², Karen L. Bell²², Daniel Marson²³, Randall Griffith²³, David Clark²³, David Geldmacher²³, John Brockington²³, Erik Roberson²³, Marissa Natelson Love²³, Hillel Grossman²⁴, Effie Mitsis²⁴, Raj C. Shah²⁵, Leyla deToledo-Morrell²⁵, Ranjan Dua²⁶, Daniel Varon²⁶, Maria T. Greig²⁶, Peggy Roberts²⁶, Marilyn Albert²⁷, Chiadi Onyike²⁷, Daniel D'Agostino²⁷, Stephanie Kielb²⁷, James E. Galvin²⁸, Brittany Cerbone²⁸, Christina A. Michel²⁸, Dana M. Pogorelec²⁸, Henry Rusinek²⁸, Mory J. de Leon²⁸, Lidia Glodzik²⁸, Susan De Santi²⁸, P. Murali Doraiswamy²⁹, Jeffrey R. Petrella²⁹, Salvador Borges-Neto²⁹, Terence Z. Wong²⁹, Edward Coleman²⁹, Charles D. Smith³⁰, Greg Jicha³⁰, Peter Hardy³⁰, Partha Sinha³⁰, Elizabeth Oates³⁰, Gary Conrad³⁰, Anton P. Porsteinsson³¹, Bonnie S. Goldstein³¹, Kim Martin³¹, Kelly M. Makino³¹, M. Saleem Ismail³¹, Connie Brand³¹, Ruth A. Mulnard³², Gaby Thai³², Catherine Mc-Adams-Ortiz³², Kyle Womack³³, Dana Mathews³³, Mary Qui-ceno³³, Allan I. Levey³⁴, James J. Lah³⁴, Janet S. Cellar³⁴, Jeffrey M. Burns³⁵, Russell H. Swerdlow³⁵, William M. Brooks³⁵, Liana Apostolova³⁶, Kathleen Tingu-s³⁶, Ellen Woo³⁶, Daniel H. S. Silverman³⁶, Po H. Lu³⁶, George Bartzokis³⁶, Neill R. Graff-Radford³⁷, Francine Parfitt³⁷, Tracy Kendall³⁷, Heather Johnson³⁷, Christopher H. van Dyck³⁸, Richard E. Carson³⁸, Martha G. MacAvoy³⁸, Pradeep Varma³⁸, Howard Chertkow³⁹, Howard Bergman³⁹, Chris Hosein³⁹, Sandra Black⁴⁰, Bojana Stefanovic⁴⁰, Curtis Caldwell⁴⁰, Ging-Yuek Robin Hsiung⁴¹, Howard Feldman⁴¹, Benita Mudge⁴¹, Michele Assaly⁴¹, Elizabeth Finger⁴², Stephen Pasternack⁴², Irina Rachisky⁴², Dick Trost⁴², Andrew Kertes⁴², Charles Bernick⁴³, Donna Munic⁴³, Marek Marsel Mesulam⁴⁴, Kristine Lipowski⁴⁴, Sandra Weintraub⁴⁴, Borna Bonakdarpour⁴⁴, Diana Kerwin⁴⁴, Chuang-Kuo Wu⁴⁴, Nancy Johnson⁴⁴, Carl Sadowsky⁴⁵, Teresa Villena⁴⁵, Raymond Scott Turner⁴⁶, Kathleen Johnson⁴⁶, Brigid Reynolds⁴⁶, Reisa A. Sperling⁴⁷, Keith A. Johnson⁴⁷, Gad Marshall⁴⁷, Jerome Yesavage⁴⁸, Joy L. Taylor⁴⁸, Barton Lane⁴⁸, Allyson Rosen⁴⁸, Jared Tinklenberg⁴⁹, Marwan N. Sabbagh⁴⁹, Christine M. Belden⁴⁹, Sandra A. Jacobson⁴⁹, Sherye A. Sirrel⁴⁹, Neil Kowall⁵⁰, Ronald Killiany⁵⁰, Andrew E. Budson⁵⁰, Alexander Norbash⁵⁰, Patricia Lynn Johnson⁵⁰, Thomas O. Obisesan⁵¹, Saba Wolday⁵¹, Joanne Allard⁵¹, Alan Lerner⁵², PaulaOgrocki⁵², CurtisTatsuoka⁵², Parianne Fatica⁵², Evan Fletcher⁵³, Pauline Maillard⁵³, John Olichney⁵³, Charles DeCarli⁵³, Owen Carmichael⁵³, Smjta Kittur⁵⁴, Michael Borrie⁵⁵, T.-Y. Lee⁵⁵, Rob Bartha⁵⁵, Sterling Johnson⁵⁶, Sanjay Asthana⁵⁶, Cynthia M. Carlsson⁵⁶, Steven G. Potkin⁵⁷, Adrian Preda⁵⁷, Dana Nguyen⁵⁷, Vernice Bates⁵⁸, Horacio Capote⁵⁸, Michelle Rainka⁵⁸, Douglas W. Scharre⁵⁹, Maria Katakis⁵⁹, Anahita Adeli⁵⁹, Earl A. Zimmerman⁶⁰, DzintraCelmins⁶⁰, Alice D. Brown⁶⁰, Godfrey D. Pearson⁶¹, Karen Blank⁶¹, Karen Anderson⁶¹, Laura A. Flashman⁶², Marc Seltzer⁶², Mary L. Hynes⁶², Robert B. Santulli⁶², Kaycee M. Sink⁶³, Leslie Gordineer⁶³, Jeff D. Williamson⁶³, Pradeep Garg⁶³, Franklin Watkins⁶³, Brian R. Ott⁶⁴, Henry Querfurth⁶⁴, Geoffrey Tremont⁶⁴, Stephen Sallo-way⁶⁵, Paul Malloy⁶⁵, Stephen Correia⁶⁵, Howard J. Rosen⁶⁶, Bruce L. Miller⁶⁶, David Perry⁶⁶, Jacobo Mintzer⁶⁷, Kenneth Spicer⁶⁷, David Bachman⁶⁷, Nunzio Pomara⁶⁸, Raymundo Hernando⁶⁸, Antero Sarrael⁶⁸, Norman Relkin⁶⁹, Gloria Chaing⁶⁹, Michael Lin⁶⁹, Lisa Ravdin⁶⁹, Amanda Smith⁷⁰, Balebail Ashok Raj⁷⁰ & Kristin Fargher⁷⁰.

⁷Magnetic Resonance Unit at the VA Medical Center and Radiology, Medicine, Psychiatry and Neurology, University of California, San Francisco, USA. ⁸San Diego School of Medicine, University of California, California, USA. ⁹Mayo Clinic, Minnesota, USA. ¹⁰Mayo Clinic, Rochester, USA. ¹¹University of California, Berkeley, USA. ¹²University of Pennsylvania, Pennsylvania, USA. ¹³University of Southern California, California, USA. ¹⁴University of California, Davis, California, USA. ¹⁵MPH Brigham and Women's Hospital/Harvard Medical School, Massachusetts, ¹⁶Washington University St. Louis, Missouri, USA. ¹⁷Oregon Health and Science University, Oregon, USA. ¹⁸University of California-San Diego, California, USA. ¹⁹Banner Alzheimer's Institute, USA.

²⁰University of Michigan, Michigan, USA. ²¹Baylor College of Medicine, Houston, State of Texas, USA. ²²Columbia University Medical Center, South Carolina, USA. ²³University of Alabama – Birmingham, Alabama, USA. ²⁴Mount Sinai School of Medicine, New York, USA. ²⁵Rush University Medical Center, Rush University, Illinois, USA. ²⁶Wien Center, Florida, USA. ²⁷Johns Hopkins University, Maryland, USA. ²⁸New York University, NY, USA. ²⁹Duke University Medical Center, North Carolina, USA. ³⁰University of Kentucky, Kentucky, USA. ³¹University of Rochester Medical Center, NY, USA. ³²University of California, Irvine, California, USA. ³³University of Texas Southwestern Medical School, Texas, USA. ³⁴Emory University, Georgia, USA. ³⁵University of Kansas, Medical Center, Kansas, USA. ³⁶University of California, Los Angeles, California, USA. ³⁷Mayo Clinic, Jacksonville, USA. ³⁸Yale University School of Medicine, Connecticut, USA. ³⁹McGill University, Montreal-Jewish General Hospital, Canada. ⁴⁰Sunnybrook Health Sciences, Ontario, USA. ⁴¹U.B.C. Clinic for AD & Related Disorders, Canada. ⁴²Cognitive Neurology - St. Joseph's, Ontario, USA. ⁴³Cleveland Clinic Lou Ruvo Center for Brain Health, Ohio, USA. ⁴⁴Northwestern University, USA. ⁴⁵Premiere Research Inst (Palm Beach Neurology), USA. ⁴⁶Georgetown University Medical Center, Washington D. C., USA. ⁴⁷Brigham and Women's Hospital, Massachusetts, USA. ⁴⁸Stanford University, California, USA. ⁴⁹Banner Sun Health Research Institute, USA. ⁵⁰Boston University, Massachusetts, USA. ⁵¹Howard University, Washington D. C., USA. ⁵²Case Western Reserve University, Ohio, USA. ⁵³University of California, Davis – Sacramento, California, USA. ⁵⁴Neurological Care of CNY, USA. ⁵⁵Parkwood Hospital, Pennsylvania, USA. ⁵⁶University of Wisconsin, Wisconsin, USA. ⁵⁷University of California, Irvine – BIC, USA. ⁵⁸Dent Neurologic Institute, NY, USA. ⁵⁹Ohio State University, Ohio, USA. ⁶⁰Albany Medical College, NY, USA. ⁶¹Hartford Hospital, Olin Neuropsychiatry Research Center, Connecticut, USA. ⁶²DartmouthHitchcock Medical Center, New Hampshire, USA. ⁶³Wake Forest University Health Sciences, North Carolina, USA. ⁶⁴Rhode Island Hospital, state of Rhode Island, USA. ⁶⁵Butler Hospital, Providence, Rhode Island, USA. ⁶⁶University of California, San Francisco, USA. ⁶⁷Medical University South Carolina, USA. ⁶⁸Nathan Kline Institute, Orangeburg, New York, USA. ⁶⁹Cornell University, Ithaca, New York, USA. ⁷⁰USF Health Byrd Alzheimer's Institute, University of South Florida, USA.

Authors' contributions

Y-CC initiated the study, designed the analysis procedures, performed the analysis and wrote the manuscript. J-TW did the literature survey of the epistasis selected by GenEpi and wrote the discussion. C-YC is the main advisor of Y-CC, guided this research on right way and provided ideas to optimize the methods. Y-CC, M-YH, Y-AT, C-YC and Y-JO were team members of Alzheimer's disease Dream Challenge. C-YC, Y-JO, P-HH, KMG and SWY commented on the draft and revised the manuscript. All authors read and approved the final manuscript. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

Funding

This work has been supported by the Ministry of Science and Technology of Taiwan grant 108–2221-E-002-079-MY3, 105–2911-H-002-566, 105–2221-E-002-129-MY3 and 103–2627-M002–015. The funding body played no roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The application process includes acceptance of the Data Use Agreement and submission of an online application form. The application must include the investigator's institutional affiliation and the proposed uses of the ADNI data. ADNI data may not be used for commercial products or redistributed in any way.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, 10617 Taiwan. ²Taiwan AI Labs, Taipei, 10351 Taiwan. ³Department of Dermatology, National Taiwan University Hospital, Taipei, 10002 Taiwan. ⁴Department of Biomechatronics Engineering, National Taiwan University, Taipei, 10617 Taiwan. ⁵Genome and Systems biology degree program, Academia Sinica and National Taiwan University, Taipei, 10617 Taiwan. ⁶Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, 94158 California, USA. ⁷Institute for Human Genetics, University of California, San Francisco, San Francisco, 94143 California, USA.

Received: 25 July 2019 Accepted: 14 January 2020

Published online: 24 February 2020

References

- Kingsmore SF, Lindquist IE, Mudge J, Gessler DD, Beavis WD. Genome-wide association studies: progress and potential for drug discovery and development. *Nat Rev Drug Discov*. 2008;7:221–30.
- Ozaki K, et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet*. 2002;32:650–4. <https://doi.org/10.1038/ng1047>.
- Klein RJ, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005;308:385–9. <https://doi.org/10.1126/science.1109557>.
- Pinero J, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*. 2017;45:D833–9. <https://doi.org/10.1093/nar/gkw943>.
- McCarthy MI, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*. 2008;9:356–69. <https://doi.org/10.1038/nrg2344>.
- Eichler EE, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*. 2010;11:446–50. <https://doi.org/10.1038/nrg2809>.
- Manolio TA, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461:747–53. <https://doi.org/10.1038/nature08494>.
- Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. *Nat Rev Genet*. 2009;10:241–51. <https://doi.org/10.1038/nrg2554>.
- Shriner D, Vaughan LK, Padilla MA, Tiwari HK. Problems with genome-wide association studies. *Science*. 2007;316:1840–2. <https://doi.org/10.1126/science.316.5833.1840c>.
- Carlborg O, Haley CS. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet*. 2004;5:618–25. <https://doi.org/10.1038/nrg1407>.
- Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*. 2009;10:392–404. <https://doi.org/10.1038/nrg2579>.
- Easton DF, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*. 2007;447:1087–93. <https://doi.org/10.1038/nature05887>.
- Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*. 2008;40:695–701. <https://doi.org/10.1038/ng.f.136>.
- Hindorf LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009;106:9362–7. <https://doi.org/10.1073/pnas.0903103106>.
- Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*. 2010;26:445–55. <https://doi.org/10.1093/bioinformatics/btp713>.
- Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*. 2005;37:413–7. <https://doi.org/10.1038/ng1537>.
- Wei WH, Hemani G, Haley CS. Detecting epistasis in human complex traits. *Nat Rev Genet*. 2014;15:722–33. <https://doi.org/10.1038/nrg3747>.
- Schubach T, Xenarios I, Bergmann S, Kapur K. FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*. 2010;26:1468–9. <https://doi.org/10.1093/bioinformatics/btq147>.

19. Wan X, et al. BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet.* 2010;87:325–40. <https://doi.org/10.1016/j.ajhg.2010.07.021>.
20. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559–75. <https://doi.org/10.1086/519795>.
21. Chang CC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7. <https://doi.org/10.1186/s13742-015-0047-8>.
22. Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension. *Ann Med.* 2002;34:88–95.
23. Yang P, Ho JW, Yang YH, Zhou BB. Gene-gene interaction filtering with ensemble of filters. *BMC Bioinformatics.* 2011;12 Suppl 1:S10. <https://doi.org/10.1186/1471-2105-12-S1-S10>.
24. Bureau A, et al. Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol.* 2005;28:171–82. <https://doi.org/10.1002/gepi.20041>.
25. Schwarz DF, Konig IR, Ziegler A. On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics.* 2010;26:1752–8. <https://doi.org/10.1093/bioinformatics/btq257>.
26. Wan X, et al. MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study. *BMC Bioinformatics.* 2009;10:13. <https://doi.org/10.1186/1471-2105-10-13>.
27. Murk W, DeWan AT. Exhaustive genome-wide search for SNP-SNP interactions across 10 human diseases. *G3 (Bethesda).* 2016;6:2043–50. <https://doi.org/10.1534/g3.116.028563>.
28. Allen GL, et al. Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease. *Alzheimers Dement.* 2016;12:645–53. <https://doi.org/10.1016/j.jalz.2016.02.006>.
29. Ma L, Clark AG, Keinan A. Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genet.* 2013;9:e1003321. <https://doi.org/10.1371/journal.pgen.1003321>.
30. Oh S, et al. A novel method to identify high order gene-gene interactions in genome-wide association studies: gene-based MDR. *BMC Bioinformatics.* 2012;13 Suppl 9:S5. <https://doi.org/10.1186/1471-2105-13-S9-S5>.
31. Li S, Cui Y. Gene-centric gene-gene interaction: a model-based kernel machine method. *Ann Appl Stat.* 2012;6:1134–61.
32. Wu X, et al. A novel statistic for genome-wide interaction analysis. *PLoS Genet.* 2010;6:e1001131. <https://doi.org/10.1371/journal.pgen.1001131>.
33. Kent WJ, et al. The human genome browser at UCSC. *Genome Res.* 2002;12:996–1006. <https://doi.org/10.1101/gr.229102> Article published online before print in May 2002.
34. Rosenbloom KR, et al. The UCSC genome browser database: 2015 update. *Nucleic Acids Res.* 2015;43:D670–81. <https://doi.org/10.1093/nar/gku1177>.
35. Lewontin R. C. the interaction of selection and linkage. I. General considerations; heterotic models. *Genetics.* 1964;49:49–67.
36. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33:1–22.
37. Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc Series B Stat Methodology.* 2010;72:417–73.
38. Urbanowicz RJ, Meeker M, LaCava W, Olson RS, Moore JH. Relief-based feature selection: introduction and review arXiv preprint arXiv:1711.08421; 2017.
39. Urbanowicz RJ, et al. GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Min.* 2012;5:16. <https://doi.org/10.1186/1756-0381-5-16>.
40. Consortium, G. T. The genotype-tissue expression (GTEx) project. *Nat Genet.* 2013;45:580–5. <https://doi.org/10.1038/ng.2653>.
41. Saura CA, Parra-Damas A, Enriquez-Barreto L. Gene expression parallels synaptic excitability and plasticity changes in Alzheimer's disease. *Front Cell Neurosci.* 2015;9:318. <https://doi.org/10.3389/fncel.2015.00318>.
42. Uhrig M, et al. New Alzheimer amyloid beta responsive genes identified in human neuroblastoma cells by hierarchical clustering. *PLoS One.* 2009;4:e6779. <https://doi.org/10.1371/journal.pone.0006779>.
43. Pietrzak M, Rempala G, Nelson PT, Zheng JJ, Hetman M. Epigenetic silencing of nucleolar rRNA genes in Alzheimer's disease. *PLoS One.* 2011;6:e22585. <https://doi.org/10.1371/journal.pone.0022585>.
44. Lambert JC, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet.* 2013;45:1452–8. <https://doi.org/10.1038/ng.2802>.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

