GenEpi: Gene-based Epistasis Discovery Using Machine Learning

Yu-Chuan Chang[1], June-Tai Wu[2], Ming-Yi Hong[3], Yi-An Tung[4], Ping-Han Hsieh[3], Sook Wah Yee[5], Kathleen M. Giacomini[5,6], Yen-Jen Oyang[1], and Chien-Yu Chen[3,*], for the Alzheimer's Disease Neuroimaging Initiative†

[1]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 10617, Taiwan. [2]Graduate Institute of Molecular Medicine, School of Medicine, National Taiwan University, Taiwan. [3]Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei 10617, Taiwan. [4]Genome and Systems biology degree program, Academia Sinica and National Taiwan University, Taipei 10617, Taiwan. [5]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California, USA. [6]Institute for Human Genetics, University of California, San Francisco, San Francisco, California, USA.

*To whom correspondence should be addressed.

†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.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 complete listing of ADNI investigators can be found at:
https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Chien-Yu Chen
Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei 10617, Taiwan.
Phone: +886-2-3366-5334, E-mail: chienyuchen@ntu.edu.tw

Abstract

Motivation: Genome-wide association (GWA) studies provide a powerful means to identify associations between genetic variants and phenotypes. However, GWA techniques for detecting epistasis, the interactions between genetic variants associated with phenotypes, are still limited. We believe that developing an efficient and effective GWA 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). In this regard, this study presents GenEpi, a computational package to uncover epistasis associated with phenotypes by the proposed machine learning approach, and illustrates the application of GenEpi on predicting the diagnosis and the progression of AD.

Methods: The key concept of GenEpi is a feature extraction process based on gene structures. Since a gene is the min-imal physical and functional unit of heredity, GenEpi considers a gene as a unit to retrieve genetic variants as features. GenEpi adopts combinatorial encoding when producing features and models them by randomized regression algorithm. Features are first modeled using machine learning approach to detect the epistasis within a single gene. The selected features for each gene are then consider together in order to identify cross-gene epistasis. This study uses AD as an example to reveal the capability of GenEpi in finding

disease-related variants and variant interactions that show both bio-logical meanings and predictive power.

Results: We compared GenEpi with commonly used algorithms for detecting epistasis, including FastEpistasis, BOOST and ReliefF. Our results suggest that the epistasis selected by GenEpi has the best predictive power for two major phenotypes in the AD dataset: diagnosis of AD and disease progression. For diagnosis, the proposed model of predicting AD contains three clinical features (Age, Gender and Education) and 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 0.83 and 0.81, respectively. On the other hand, for predicting progression, the proposed model contains eight clinical features (Age, Gender, Education, Cognitively Normal (CN), Early Mild Cognitive Impairment (MCI), Late MCI, AD, and MMSE at baseline) and four genetic features, including seven SNPs from six genes, where all of the four genetic features are cross-gene epistasis with significant p-values ($< 10^{-11}$). The average of Pearson and Spearman correlation of 2-fold CV and LOO CV are 0.52 and 0.53, respectively. The results on 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.

Availability: GenEpi is an open-source python package and available free of charge only for non-commercial users. The package can be downloaded from https://github.com/Chester75321/GenEpi, and has also been published on The Python Package Index.

Introduction

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 three 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 2017, there have been more than hundreds of thousands individuals being studied in typical GWAS protocols, and over 72,870 variant-disease associations between 46,589 SNPs and 6,356 phenotypes have been discovered [4]. These studies demonstrated the potential of GWAS to identify genetic variants associated with many categories of phenotypes, including risk for

85 diseases such as various cancers, and variation in therapeutic and adverse response to drugs.
86 However, the success of univariate GWAS is limited to monogenic phenotypes (e.g. Mendelian
87 diseases). The impact of variant interactions, also known as epistasis on the formation of
88 diseases [5] is often underestimated in traditional GWAS analysis [6-8].
89 Researchers have found that GWAS has limitation in identifying the association in complex
90 diseases [9, 10]. Easton et al. suggested that a number of susceptible loci identified by GWAS
91 usually have very small effect sizes [11]. A major disadvantage of GWAS is that it considers
92 only one genetic variant at a time, and ignores underlying epistasis of variants that might have
93 stronger associations [12]. Studies have also demonstrated that the existence of epistasis is an
94 important factor contributing to phenotypes, especially in complex diseases such as
95 hypertension, diabetes and obesity [10]. Therefore, developing analytical methods to identify
96 epistasis efficiently is critical to understanding the genetic factors that contribute to human traits
97 [8, 13], and has attracted a wide range of research interests in recent years [7, 14].
98 Two main challenges to discover epistasis are computational complexity and statistical power
99 [15]. The first challenge results from the curse of dimensionality. When more genetic variants
100 are considered, the number of interactions increases exponentially. Based on the specification
101 of a major commercial technology, Illumina Arrays, a whole-genome array can investigate over

102 4 million markers per sample simultaneously. In order to evaluate the pairwise interactions from
103 this microarray, about 8×10^{12} statistical tests need to be processed. Even though Marchini et al.
104 have demonstrated that pairwise interactions of 3×10^{5} loci is computationally possible with
105 currently available computational resources, it still remains challenging when the Illumina
106 Arrays are considered [16]. The second challenge is the issue of statistical power. Since a huge
107 number of statistical tests are conducted on a limited sample size with high-dimensional
108 interactions, many false positives arise by random chances. In this regard, new methods have
109 been developed [10, 17]. Statistical approaches include FastEpistasis [18] and BOOST [19];
110 the later one has been included in a well-known GWAS software called PLINK [20, 21].
111 Machine learning approaches such as Multifactor Dimensionality Reduction [22], ReliefF [23],
112 random forest-like algorithms [24-26] and other methodologies have also been developed for
113 detecting epistasis [17].
114 Since the biological experiments used to validate these methodologies are still in demand, there
115 are no standard analysis methods for epistasis despite the rapid improvement in computational
116 performance. In 2016, Murk used FastEpistasis and BOOST to search SNP-SNP interactions
117 on a huge dataset called Genetic Epidemiology Research on Adult Health and Aging (GERA)
118 that included 78,486 subjects, but still failed to detect a significant and replicable interaction

119 after exhaustively searching through 45 billion possible interactions for 10 complex diseases of
120 interest [27]. Alzheimer's disease (AD) is one of these complex diseases and its pathogenesis,
121 which clearly has a genetic basis, is still ill-defined. In 2016, Sage Bionetworks held a
122 competition called The Dialogue for Reverse Engineering Assessments and Methods Challenge
123 (DREAM Challenge) for AD, which tried to use crowdsourcing to assess the capability of
124 current computational methods to predict the change in cognitive examination based on genetic
125 data. However, no significant contribution of genetic features except the *APOE* haplotype to
126 the predictive performance was observed by any competition teams [28]. To address this
127 problem, this study presents GenEpi, a package to reveal epistasis related to the phenotype
128 using machine learning and introduces the application of GenEpi on predicting the diagnosis
129 and the progression of AD.
130

131 **Methods**

132 The architecture of GenEpi is shown in Fig. 1. GenEpi is designed to group SNPs by gene
133 boundaries. GenEpi first considers the genetic variants in a gene as features, which have a
134 higher chance to interact with each other and to influence molecular functions. The idea of
135 within-gene epistasis analysis followed by cross-gene analysis is not new, which has been

136 proven successful in previous studies [29-32]. Differently, GenEpi adopts combinatorial
137 encoding when producing features and models them by randomized regression algorithm,
138 which will be explained in Section 2.2. In the first stage (STAGE 1) of GenEpi, the genotype
139 features from each single gene will be combinatorically encoded and modeled independently
140 by randomized regression algorithm. In this way, we can estimate the prediction performance
141 of each gene and detect the within-gene epistasis with a low false positive rate. In the second
142 stage (STAGE 2), both of the individual SNP and the within-gene epistasis features selected by
143 STAGE 1 are pooled together to generate cross-gene epistasis features. Finally, the genotype
144 features after feature selection are combined with environmental factors such as clinical
145 features to build the final prediction models. In addition to the main procedures, two pre-
146 processing steps are also implemented in GenEpi: retrieving the gene information from public
147 databases and reducing the gene information from public databases and reducing the
148 dimensionality of the features using linkage disequilibrium (LD). In the end, we released a
149 Python package that implements GenEpi (steps are illustrated in Supplementary Fig. S1). The
150 details of these steps and the GenEpi method will be described in the following sections.
151
152 **University of California Santa Cruz (UCSC) database**

153 To obtain the gene information such as official gene symbols and genomic coordinates, we
154 retrieved kgXref and knownGene data table from the UCSC human genome annotation
155 database [33, 34]. The version of the database we used is the Feb. 2009 assembly of the human
156 genome hg19, GRCh37 Genome Reference Consortium Human Reference 37. The two data
157 tables were merged in order to generate a local database containing the gene symbols as well
158 as the genomic coordinates of each gene. The in-house script we built could update this local
159 database automatically. It is noted that there are many different categories of genes in the
160 RefSeq database. In this study, we only focused on the mRNA and non-coding RNA (22,376
161 genes in total). The selected transcripts were projected on the genomic coordinates and the
162 coordinates of corresponding genes were determined based on the leftmost and rightmost
163 positions of the corresponding transcripts. Moreover, to discover the factors that might affect
164 the transcription of genes, we also retained the promoter region of each gene. In genetics, the
165 promoter region is a segment of DNA that initiates the transcription of a particular gene.
166 Promoters are located near the transcription start sites of genes, on the upstream of the same
167 DNA strand (towards the 5' region of the sense strand of the transcript). In general, a promoter
168 region can be 100-1000 base pairs long. In this study, we extracted 1,000 nucleotides on the
169 upstream of the starting position of each gene as the promoter region.

170

171 **Estimation of linkage disequilibrium**

172 In GWAS datasets, a SNP often exhibits high dependency with its nearby SNPs because of
173 linkage disequilibrium (LD). In the practical implantation, we prefer to group these dependent
174 features to reduce the dimension of features. In other words, we can take the advantages of LD
175 to reduce the dimensionality of SNP features. In this regard, we adopted the same approach
176 developed by Lewontin [35] to estimate LD. We used $D' > 0.9$ and $r^2 > 0.9$ as the criteria to group
177 highly dependent SNP features as blocks. In each block, we chose the features with the largest
178 minor allele frequency to represent other features in the same block.

179

180 **Discovery of within-gene epistasis**

181 The main objective of the first stage in GenEpi is to select candidate features from each gene.
182 In order to extract SNP features for a gene, we used the start and end positions of each gene
183 from the local UCSC database to split the SNP features after dimension reduction. Since there
184 are 22,376 genes in the UCSC database, we obtained 22,376 subsets of the SNP features. In
185 each subset, A SNP feature with the alleles 'A' and 'a' could have three possible genotypes, AA,
186 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 following equation, $\boldsymbol{\alpha_1 SNP_1} + \boldsymbol{\alpha_2 SNP_2}$ stand for the additive interactions and $\boldsymbol{\alpha_{int(1,2)} SNP_1 \otimes SNP_2}$ represents the synergistic interactions that contain nine terms shown in equation.

$$y \sim \alpha_0 + \sum_{m \in \{AA,Aa,aa\}} \alpha_{1,m} SNP_{1,m} + \sum_{m \in \{AA,Aa,aa\}} \alpha_{2,m} SNP_{2,m} + \sum_{m,n \in \{AA,Aa,aa\}} \alpha_{1,m,2,n} SNP_{1,m} SNP_{2,n}$$

$$= \alpha_0 + \alpha_1 SNP_1 + \alpha_2 SNP_2 + \alpha_{int(1,2)} SNP_1 \otimes SNP_2$$

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 genotype among the total samples in the dataset. The second criterion is regarding the association between the feature and the phenotype. If the phenotype is dichotomous or categorical, the p-value of χ^2 test should be smaller than 0.01. If the phenotype is a real number, we used F-test to estimate the correlation 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 following equation, which be termed combinatorial encoding.

$$y \sim \alpha_0 + \sum_{i=1}^{k} \alpha_i SNP_i + \sum_{i=1}^{k} \sum_{j=1 \cap j \neq i}^{k} \alpha_{int(i,j)} SNP_i \otimes SNP_j$$

We conducted Randomized Lasso and L1-penalized Logistic regression for modeling each gene. The first advantage of them is that the L1-penalized model prefers solutions with a smaller number of features, which effectively reduces the effect of feature dependency. The second advantage is derived from randomization techniques [36], which can reduce the false positive rate. Randomized regression works by resampling the training set and remodeling in multiple times, which are able to pick out the features that are repeatedly selected across randomization. In this study, we executed this randomization 500 times, and the features selected by randomization 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 and then 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 or F-test to

221 control the quality of features and to avoid overfitting. Nevertheless, the p-value of each feature
222 in this stage should be smaller than 10^{-5}. All of the features from different genes would be
223 merged for modeling cross-gene epistasis. We conducted Lasso and L1-penalized Logistic
224 regression for modeling, and the randomization techniques were used once again to select the
225 final genotype feature set. Since the phenotype may also be affected by environmental factors,
226 after determining the final set of genotype features, we included the environmental factors such
227 as clinical assessments for constructing the final model. To obtain the final model, we used
228 random forests with 1,000 decision trees as the ensemble algorithm. Subsequently, the final
229 model was evaluated through a process called double cross validation (CV). In the external loop
230 of double CV, all the instances were divided into two subsets to serve as training and
231 independent test sets. In this study, we used 2-fold CV and leave-one-out CV (LOO CV) in
232 external loop for evaluation. In the internal loop, we also used 2-fold CV for model selection.
233

234 **Materials**

235 This study applied GenEpi on an AD cohort, which was used in Alzheimer's disease Dream
236 Challenge [28], In total, the cohort consists of 767 participants, who were healthy elderly, mild
237 cognitive impairment (MCI) and AD patients from the Alzheimer's Disease Neuroimaging

238 Initiative (ADNI) database. The 767 ADNI participants consist of 241 cognitively normal (CN),
239 130 Early MCI (EMCI), 273 Late MCI (LMCI) and 123 AD participants. We adopted two types
240 of features in this study: clinical and genetic data. The clinical assessment includes the 'Mini-
241 Mental State Examination (MMSE) at baseline' and the '24-month follow-up MMSE'. Other
242 clinical measures included 'diagnosis at baseline', 'age at baseline' (Age), 'years of education
243 at baseline' (Education), 'gender' (Gender), and '*APOE* genotype'. All the genetic data has
244 been pre-processed by the organizers that held the challenge [28]. The genetic data were
245 genotyped using the Illumina Human610-Quad BeadChip and Illumina HumanOmniExpress
246 BeadChip. The multidimensional scaling analysis was applied by PLINK using HAPMAP3 to
247 ensure that samples are within the cluster of European populations. Subsequently, the data were
248 imputed according to the 1,000 genome haplotypes. After imputation, there were 12,809,667
249 genotype features in total. Because GenEpi can be applied to solve both classification and
250 regression problems, this study illustrated both applications of GenEpi on the ADNI cohort. In
251 predicting the diagnosis of AD, we used 364 participants, of which the clinical diagnosis are
252 CN or AD, to predict which samples are control subjects or the AD patients. For predicting the
253 progression of AD, we used the MMSE difference between '24-month follow-up MMSE' and
254 'MMSE at baseline' as the predicting targets.

Experimental Results

This section includes four parts. In the first part, we applied GenEpi and other algorithms for detecting epistasis, including FastEpistasis [18], BOOST[19] and ReliefF [23, 37] on simulation data for vali dation and efficiency 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 × (precision × recall) / (precision + recall)). In the third part, we used the same dataset to predict the progression of AD, which is being defined as the MMSE difference between the baseline and the follow-up after 24 months, and the performance was assessed using Pearson correlation and Spearman correlation between the predicting results and the MMSE differences. In the final part, we compared GenEpi with other algorithms we mentioned in first part on the ADNI dataset in terms of computing time and prediction performance in real data.

Experiments on simulation data

In this section, we applied different algorithms on simulation data for validation and efficiency comparison. All of the simulation datasets are generated by the simulator GAMETES [38],

272 which is publicly accessible on the web site

273 https://popmodels.cancercontrol.cancer.gov/gsr/packages/gametes/. We designed two types of

274 simulation datasets: basic and complex models. The 'Model 1', 'Model 2' and 'Model 3' are

275 simulation datasets with the basic model, which means that each dataset only contained one

276 epistasis consisting of a SNP pair. All of the basic-model datasets are in the same setting as

277 follows: #individuals = 2000, case/control ratio is one, #SNPs = 100, #replicates = 100, minor

278 allele frequency of target SNPs = 0.2, heritability = 0.2. The complex model means one dataset

279 is allowed to contain multiple epistasis from different SNP pairs. Thus, the 'Combined Model

280 1+2+3' is a complex-model dataset containing three epistasis from the previous three basic

281 models. Fig. 2 shows the results of these four simulation datasets. Fig. 2(a) shows that the

282 ranking of the target epistasis reported by GenEpi in 300 replicates of the three basic-model

283 datasets are always ranked the first. For FastEpistasis and BOOST, the medians of the ranking

284 of the target epistasis are one but there existed some outliers, which means that they will obtain

285 many false positive epistasis for the sake of colleting the target epistasis. For the result of the

286 complex-model dataset in Fig. 2(b), the advantage of GenEpi over other algorithms is more

287 obvious. Since the Python package called scikit-rebate [37] that we used for implementing

288 ReliefF would not report the scores for each epistasis but instead the importance of each SNPs,

289 we listed the medians of the ranking for each SNP in Table 1. Table 1 reveals that ReliefF can
290 detect target epistasis in the basic models. However, GenEpi still outperforms ReliefF in the
291 complex model.
292 We believed that the advantage of GenEpi is derived from the combinatorial encoding of the
293 genotype features and the randomization techniques. In contrast with other statistical algorithms
294 such as FastEpisasis and BOOST, which only evaluate the epistasis within one SNP pair in one
295 time, GenEpi looks for the epistasis among all of the SNP combinations. Moreover, the false
296 positives among the epistasis can be filtered out by resampling and remodeling the dataset in
297 multiple times. Therefore, we applied both randomized regression and original regression on
298 the complex-model dataset to compare the false positive rates as shown in Fig. 3. It shows that
299 randomized regression can reduce the false positive rate indeed.
300
301 **Classification – Classifying AD patients**
302 In predicting control subjects or AD patients, we applied GenEpi on the 364 samples with CN
303 (as control) or AD. After dimensionality reduction, 12,102,888 out of the 12,809,667 SNPs in
304 the ADNI dataset were retained, and 4,916,249 of them are located in 20,206 genes (Table 2).
305 In the step 4 of selecting epistasis, there are 34,689 genetic features selected and 765 of them

306 are single SNP features, while the other 33,924 are epistasis features within genes. The final
307 model contained 3 clinical features (Age, Gender, and Education) and 14 genetic features,
308 including 24 SNPs from 12 genes. These features contained two single SNP features, 11 within-
309 gene epistasis features and one cross-gene epistasis feature.

310 To validate the effectiveness of the genetic features extracted by GenEpi, we compared different
311 subsets of the features in Table 3. The performance of the model considered only clinical
312 features outperform that considered only genetic features. When we combined both kind of
313 features, the performance improved, showing that added genetic features is effective. To reduce
314 the redundancy, we removed '*APOE* genotype', which can be replaced by the genetic features
315 selected by GenEpi. We also removed 'MMSE at baseline', which is a clinical assessment to
316 measure the cognitive impairment. We argued that it does not make sense to use 'MMSE at
317 baseline' as a feature to predict the status of AD at the moment of diagnosis. The last two
318 columns of Table 3 showed that the model that used both genetic and clinical features without
319 *APOE* genotype and 'MMSE at baseline' is significantly better than the one that considered
320 only clinical features.

321 We listed the statistical significance of the selected genetic features in Table 4. The first column
322 lists each feature by its RSID (Rapid Stain Identification Series) and the genotype (denoted as

323 RSID_genotype), the pairwise epistasis features are represented using two SNPs. The last
324 column describes the genes where the SNPs are located according to the genomic coordinates.
325 We used a star sign to denote the epistasis between genes. Here, only the feature
326 (rs3130614_BB, rs41276317_AB) is cross-gene epistasis (for MICB and TOB2). The weights
327 in the second column were extracted from the linear model we defined in Section 2.4. The signs
328 of the weights indicate if a feature is a causal or protective genotype. On the other hand, the
329 odds ratio significantly away from 1 also indicates whether the features are potential causal or
330 protective genotypes. Since low genotype frequency may cause unreliable odds ratios, we also
331 listed this information in the table. Lastly, the p-value of the χ^2 test showed that these features
332 are significantly associated with the phenotype. Figure 5 are the bar plots for the features
333 selected by GenEpi shown in Table 4. We took one single-SNP and the cross-gene epistasis
334 features as examples. Since each SNP pair has nine SNP combinations, we used the color 'light
335 blue' to highlight the features selected by GebEpi. We considerd that a good feature can be
336 judged by two conditions. The first one is that the difference between the amount of case and
337 control subjects should be large, which suggests that the case and control subjects can be
338 distinguished by this feature. The second one is that the genotype frequency of this feature
339 cannot be too small, which suggests that the occurrences of this feature are not only supported

340 by a small number of samples. Fig. 5(a) reveals that rs429358_AA is a good feature, because
341 the number of samples in the light blue cell is about 60 percent of the whole dataset and the
342 proportion of control versus case is 4.425:1. This feature is actually a single SNP feature located
343 on *APOE*. Moreover, it is a well-known SNP for determining the allele type of *APOE*. Fig. 5(b)
344 showed all combinations of the SNPs rs3130614 and rs41276317. Even though the number of
345 samples in the light blue cell is not particularly large, it can still be considered as a good feature
346 with following explanation. By examining the cells in this figure, two of them are empty cells,
347 one of them has only one sample, and another five cells have more control subjects than cases
348 in the cells. Since the proportion of control subjects and cases of this dataset is 2:1, the samples
349 seem to be randomly distributed into these five cells. In contrast, the proportion of the control
350 subjects and cases is 1:8 in the light blue cell in the opposite, which indicates the SNP pair with
351 such combination can be considered as a potential causal genetic variant for AD. These results
352 show that GenEpi has the ability to identify good features.
353
354 **Regression – Predicting the progression of AD**
355 In predicting the progression of AD, we applied GenEpi on all of the samples from ADNI. For
356 this task, the 'diagnosis at baseline' and 'MMSE at baseline' were also adopted as clinical

357 features. After LD dimension reduction, the original 12,809,667 SNPs decreased to 12,507,343,
358 and 5,086,639 of them are located on 20,206 genes (Table 2). After selecting epistasis by Step
359 4, there are 56,427 genetic features left, and 1302 of them are single SNP features, while the
360 other 55,125 are epistasis features in a single gene. The proposed model contained eight clinical
361 features (Age, Gender, Education, CN, EMCI, LMCI, AD, and MMSE at baseline) and four
362 genetic features, which included seven SNPs from six genes, and all of the four genetic features
363 are cross-gene epistasis. We compared different subsets of features in Table 5. The performance
364 of the model that only included genetic features is lower than the one that only included the
365 clinical features. However, when we combined both of them, the performance improved, similar
366 to the situation observed in the task of classification. It shows that the genetic features are
367 effective. Again, the *APOE* genotype in the clinical feature set is a genetic feature in essence.
368 So, we removed it from the list of clinical features. The last two columns showed that even
369 though we removed the *APOE* genotype, the model with genetic features are still better than
370 the model without the genetic features. The weights in Table 6 were extracted from the linear
371 model we described in Section 2.4. The signs of the weights indicate whether the feature is a
372 causal or protective gene. We applied student's t-test on the samples with the combination of
373 the selected genetic features (Fig. 6), and all of the p-values of t-test are significant ($< 10^{-11}$).

These results show that GenEpi has the ability to identify good features in the regression task, too.

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 contains 32 CPUs and 256 GB RAM. Since the second generation of PLINK has imported FastEpistasis and BOOST, we used PLINK to test these two algorithms. For ReliefF, we employed a Python package called scikit-rebate [37] 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 7). However, since FastEpistasis only supports the classification trial in PLINK, it was skipped in the task of regression. 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 and 5,086,639 in classification and regression tasks, respectively. BOOST took the same subsets of SNPs as GenEpi (Table 7). When taking the same subset of SNPs as GenEpi and BOOST, ReliefF still caused memory errors. Therefore,

391 we used the subsets of SNPs that selected by STAGE 1 of GenEpi as the input of ReliefF, which
392 are 33,868 and 69,842 in classification and regression tasks, respectively. We selected the top
393 15, 30, 45 and 60 rankings from the results of these algorithms for comparing the prediction
394 performance. Since different algorithms adopted different strategies to estimate the importance
395 of the feature, we adopted different formulas defined in Section 2 to generate features for
396 prediction. FastEpistasis and BOOST used χ^2 statistics to rank the SNP pairs. For these two
397 algorithms, we put the top SNP pairs into Equation 2 to generate pairwise interactions for each
398 pair. ReliefF reported significance scores for each SNP, but we cannot find a way to know the
399 relations between these SNPs. Therefore, we used Equation 3 to generate the interactions for
400 all the combinations of SNP pairs. Finally, we used logistic and lasso regression to build the
401 models for classifying AD patients and predicting the progression of AD for comparison. Table
402 7 show that GenEpi is the most efficient method, which only cost 9.95 and 28.38 CPU-days in
403 classification and regression tasks, respectively. In both classification or regression tasks,
404 GenEpi had the best prediction performance despite the fact that GenEpi only uses the subset
405 of SNPs from the final model. GenEpi shows that the time needed for identifying epistasis can
406 be drastically reduced, without compromise to the performance. We provided the ROC curves
407 for the classification task in Fig. 7, and it shows that GenEpi achieved the best performance in

408 both training and double 2-fold CV procedures, of which the area under the curve (AUC) are

409 0.95 and 0.85, respectively.

410

411 **Discussion**

412 The results in the previous section revealed the power of GenEpi to identify phenotype-

413 associated epistasis efficiently. GenEpi selected 14 features from 12 genes to categorize patients

414 with AD and four features from six genes to predict the progression of AD. Since AD is a

415 chronic neurodegenerative disease, our findings would be supported if the gene identified by

416 GenEpi are expressed in brains. We downloaded the datasets of the GTEx Project [39] to inspect

417 the gene expression of these genes in different tissues, as shown in Figure. 6. Among the 18

418 genes selected by GenEpi, 13 have high expression level in the brain tissues. In addition, five

419 genes, *VSNL1*, *PRKAG2-AS1*, *SYT6*, *CLEC1A* and *LINC00299*, have a similar expression

420 pattern with *APOE*. These 18 genes are categorized as cross-gene epistasis, single-gene

421 epistasis and single-SNP features based on the feature types selected by GenEpi. The detail

422 discussion about these three types of feature will be describe in following sections.

423

424 **Cross-gene epistasis**

425 GenEpi picked out five cross-gene epistasis in total, which include *CLEC1A * GPR142,*

426 *CLEC1A * PRKAG2-AS1, IFT20 * RPL32, MFSD6L * PRKAG2-AS1* and *MICB * TOB2*. We

427 found several direct and indirect evidence to demonstrate that these interactions might have

428 true association with AD. Liddelow et al. found that neurotoxic reactive astrocytes are induced

429 by activated microglia [40] and thus are implicated in AD [41]. The expression of *CLEC1A*, a

430 phagocytic receptor gene, increased in aged microglia, indicative of a pro-inflammatory status

431 [42]. Based on HaploReg [43], the enhancer of *GPR142* is DNAse sensitive in primary

432 astrocytes. Together evidence suggests that *CLEC1A* and *GPR142* are expressed in two

433 inflammatory cells strongly related in AD. Li et al. found that *PRKAG2-AS1* and *PRKAG2* were

434 differentially regulated [44], suggesting a possible negative regulatory role of *PRKAG2-AS1* on

435 the *PRKAG2* mRNA level. The *PRKAG2*, AMP-activated protein kinase subunit gamma-2, has

436 a polymorphism associated with cognitive impairment as well as diabetes in old age [45].

437 AMPK activation is also related to cognitive impairment in old age, neuroinflammation, and

438 neurodegeneration [46]. Together, the epistasis between *PRKAG2-AS1* and *CLEC1A* raises an

439 interesting possibility that *PRKAG2-AS1* may functionally interact with *CLEC1A* through

440 AMPK regulated neuroinflammation.

441 We did not find direct link between *RPL32* and AD, but by the data presented in HaploReg

442 there is another gene called *IQSEC1* that is located in the same LD block with *RPL32*. *IQSEC1*
443 is expressed in neuron and physically binds to *APOE*, but the implication of this interaction in
444 AD is not clear. *IFT20* is an essential component of primary cilia that appears to be important
445 in the survival of dentate granule cell, which plays a key role in memory formation and related
446 to AD [47]. These studies showed that *IFT20* and *IQSEC1* have been independently implicated
447 in AD and our results suggest they might together maintain an intact primary cilia that functions
448 critically in AD. MIC proteins are induced on the cell surface by stress and are ligands of a
449 common activatory natural killer-cell receptor (*NKG2D*) [48]. They thus have roles in
450 antimicrobial defense, tumor suppression and autoimmune-like diseases [49]. Quiroga et al. has
451 found that the human MHC class I chain-related genes (*MICA* and *MICB*) are located within
452 the HLA region that has stratified association with AD in the cohort of the Oxford Project to
453 Investigate Memory and Ageing (OPTIMA) even though the association did not survive the
454 correction for multiple testing [50]. *TOB2* has been reported as a most up-regulated gene [51]
455 by expression analysis on multiple AD expression datasets from the GEO database to identify
456 significant genes associated with electrophysiological pathways and attempted determination
457 of interconnected canonical molecular pathways. The interaction of *MICB* and *TOB2* suggests
458 a not-yet-explored biological process in AD pathogenesis. However, the significant

459 associations of these genes in genome-wide association studies associated with immune system

460 diseases and neurological diseases (see GWAS Catalog, 22561518, 29083406, 28540026),

461 warrant future investigation. Among all the cross-gene epistasis selected by GenEpi, only one

462 gene *MFSD6L* cannot find the association with AD so far.

463

464 **Single-gene epistasis**

465 There are several possible modes of action accounting for intramolecular SNP-SNP interactions

466 identified in this study. First: a synergistic regulation of transcription. For example, rs12095538

467 resides in the eQTL of STY6 and rs2774308 resides in a DNAse region of STY6. It is very

468 possible that variants in both rs12095538 and rs2774308 synergistically affect the expression

469 of STY6. SYT6 encodes protein control neurotransmitter release. It has been shown that the

470 expression of SYT6 is significantly down regulated in AD [52], supporting the causative role

471 of the inherited rs12095538 and rs2774308 haplotype in AD progression. Likewise,

472 rs56233035 and rs3678 reside in the LD blocks containing the eQTL and DNAse region of

473 CACNA1E, respectively. CACNA1E encodes a subunit of CaV1.2 calcium channel. Its

474 expression level in reactive astrocytes is associated with amyloid-β plaques formation in an AD

475 mouse model [53]. Another examples are the interactions between rs12926153 and rs12922908,

476 and rs9652600 and rs12922908 that are within DNAse regions controlling the expression of
477 CLEC16A in brain. CLEC16A encodes a lectin receptor involved in inflammation process and
478 is differentially expressed in AD brains [54]. Lastly, rs2052573 and rs34580133 reside in
479 DNAse regions within the LINC00299 and thus may synergistically control the expression of
480 this long non-coding RNA that has been implicated in brain development [55].
481 Second: a synergestic interaction between transcriptional and post-transcriptional regulation.
482 For example, rs9344977, in an intron of BACH2 containing cis-regulatory elements, may affect
483 the expression of BACH2, whereas rs56148686 in another intron of BACH2 without any cis-
484 regulatory element. We suspect that rs56148686 may interact with s9344977 through a post-
485 transcriptional mechanism such as RNA stability, RNA splicing, or microRNA binding.
486 BACH2 encodes a transcription factor involving cellular responses toward oxidative stress.
487 Interestingly, BACH2 is upregulated in cultured human neuroblastoma cells upon exposure to
488 Alzheimer Amyloid β [56]. Similar mechanisms may apply to interaction pairs rs12189429-
489 rs6881360 and rs12187423-rs6881360 that presumably regulating the abundance of
490 ADAMTS12 mRNA. The importance of ADAMTS12 in AD is supported by another
491 independent study using same ANDI dataset [57]. Third: an intramolecular SNP pairs
492 modulating the expression of two separate neighboring genes. For example, rs11675339 and

493 rs2710687 of VSNL1 appear to control the expression of GEN1 and SMC6 respectively. We
494 also found that VSNL (also known as VILIP-1) has strong evidences to AD [58-60]. GEN1 and
495 SMC6 play pivotal roles in repairing double-strand breaks of genomic DNA. Interestingly,
496 rDNA instability, a result of failures to repair double-strand breaks of genomic DNA, has been
497 implicated in AD [61]. Most of the single-gene epistasis selected by GenEpi can be explained
498 by these three possible modes and only two of the significance SNP-SNP interactions are not
499 immediately clear at this moment. For example, rs12366151 resides in the eQTL of MICALCL,
500 an AD associated gene [62] but the interacting rs10831829 resides in the intron of PARVA, a
501 gene not-yet-reported to be associated with AD. Similarly, rs2421701 and rs200512701 reside
502 in two eQTLs regulating TNKS2 expression. However, the ADP-ribose polymerase
503 functionally contributes to AD is not clear.
504
505 **Single-gene epistasis**
506 There are only two single-SNP features and both of them are located in *APOE*, which is a well-
507 known causal gene of AD, revealing that GenEpi is an effective tool to identify disease-causing
508 genes. Moreover, GenEpi successfully selected out the SNP rs429358, which determines the
509 allele type of *APOE* with rs7412.

510

511 ## Conclusion

512 While GenEpi has shown its ability to identify epistasis efficiently, it might still has the
513 following limitations. Firstly, GenEpi can only detect pairwise interactions. Second, GenEpi is
514 a memory-consuming package, which might cause memory errors when calculating the
515 epistasis of a gene containing a large number of SNPs. We recommend that the memory for
516 running GenEpi should be over 256 GB. Finally, a small sample size may lead overfitting,
517 which forces us to use strict thresholds during feature selection. In this way, GenEpi delivers a
518 high precision rate, but might suffer having false negatives. This implies different GWAS data
519 might detect different sets of true positives. In summary, the results of this study demonstrated
520 that GenEpi is a promising software package to identify causal SNPs and epistasis in GWAS,
521 and it can be further used to predict the phenotypes. With the demonstrated efficiency, GenEpi
522 is a powerful tool to explore gene-gene interactions that underlie complex diseases.
523

Tables

Table 1. The medians of the ranking of the target epistasis for ReliefF

	SNP 1	SNP2	SNP3	SNP4	SNP5	SNP6
Basic-model	1	2	1	2	1	2
Complex-model	7	8.5	9.5	11.5	11	19.5

Table 2. The summary of the number of features after feature selection in the classification and regression tasks, respectively.

Step	Classification	Regression	Unit
0 (raw data)	12,809,667	12,809,667	SNP
1	22,376	22,376	Gene
2	12,102,888	12,507,343	SNP
3	4,916,249	5,086,639	SNP
	20,206	20,206	Gene
4	44,865	69.843	SNP
	7,933	9,561	Gene
	34,689	56,427	Feature
5	24	7	SNP
	12	6	Gene
	14	4	Feature

Table 3. Comparison of different subsets of features in predicting control subjects or AD patients.

		Genetic	Clinical	Genetic + Clinical	Clinical - APOE - MMSE	Genetic + Clinical - APOE - MMSE
Training	Precision	0.9633	1.0000	1.0000	0.9658	1.0000
	Recall	0.8537	1.0000	1.0000	0.9187	1.0000
	F1 Score	0.9052	1.0000	1.0000	0.9417	1.0000
	Accuracy	0.9396	1.0000	1.0000	0.9615	1.0000
2-fold CV	Precision	0.7748	0.9280	0.9675	0.3774	0.7699
	Recall	0.6992	0.9431	0.9675	0.3252	0.7073
	F1 Score	0.7350	0.9355	0.9675	0.3493	0.7373
	Accuracy	0.8297	0.9560	0.9780	0.5907	0.8297
LOO CV	Precision	0.8100	0.9286	0.9600	0.3738	0.7596
	Recall	0.6585	0.9512	0.9756	0.3252	0.6423
	F1 Score	0.7265	0.9398	0.9677	0.3478	0.6960
	Accuracy	0.8324	0.9588	0.9780	0.5879	0.8104

'Genetic' is short for genetic features; 'Clinical' is short for clinical features. F1 Score = 2 × (Precision × Recall) / (Precision + 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.

534 **Table 4.** 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	*MICB*TOB2*
rs12095538_BB, rs2774308_AB	2.41	7.69	6.87E-07	0.0824	*SYT6*
rs12926153_AB, rs12922908_AA	1.18	4.83	6.89E-07	0.1511	*CLEC16A*
rs9652600_AB, rs12922908_AA	0.94	4.83	6.89E-07	0.1511	*CLEC16A*
rs9344977_BB, rs56148686_AB	1.94	4.32	1.14E-06	0.1813	*BACH2*
rs429358_AA	-2.01	0.17	1.73E-06	0.5962	*APOE*
rs56233035_AB, rs3678_AB	2.26	10.16	1.91E-06	0.0604	*CACNA1E*
rs11675339_AA, rs2710687_AA	2.32	3.94	3.55E-06	0.1923	*VSNL1*
rs12189429_BB, rs6881360_AA	1.36	4.34	3.65E-06	0.467	*ADAMTS12*
rs12187423_BB, rs6881360_AA	0.58	4.34	3.65E-06	0.467	*ADAMTS12*
rs10831829_BB, rs12366151_AA	3.48	9.50	4.90E-06	0.0577	*PARVA*
rs2052573_BB, rs34580133_AB	1.80	4.08	5.00E-06	0.1648	*LINC00299*
rs2421701_AB, rs200512701_AB	1.82	4.12	5.29E-06	0.1593	*TNKS2*
rs769449_AA	-1.19	0.16	8.42E-06	0.6648	*APOE*

535 The sign '*' between two gene symbols indicates cross-gene epistasis.

536

Table 5. Comparison of different subsets of features in predicting the progression of AD.

		Genetic	Clinical	Genetic + Clinical	Clinical - APOE	Genetic + Clinical - APOE
Training	Pearson	0.4641	0.9582	0.9613	0.9533	0.9600
	Spearman	0.2939	0.9277	0.9375	0.9227	0.9385
	AVG(P+S)	0.3790	0.9430	0.9494	0.9380	0.9493
2-folds CV	Pearson	0.4088	0.5014	0.5771	0.4896	0.5702
	Spearman	0.2286	0.4517	0.4915	0.4289	0.4776
	AVG(P+S)	0.3187	0.4765	0.5343	0.4593	0.5239
LOO CV	Pearson	0.4123	0.4922	0.5806	0.4741	0.5732
	Spearman	-0.1032	0.4441	0.4960	0.4197	0.4854
	AVG(P+S)	0.1545	0.4682	0.5383	0.4469	0.5293

'Genetic' is short for genetic features; 'Clinical' is short for clinical features; 'Pearson' means Pearson's Correlation; 'Spearman' means Spearman's Correlation; 'AVG(P+S)' means the average of Pearson and Spearman correlations; '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.

543 **Table 6.** The statistical significance of genetic features selected by GenEpi in predicting the progression of AD.

Selected SNPs (RSID)	Weight	T-test P-value	Genotype Frequency	Gene
rs2306894_BB, rs11653674_BB	-1.2045	5.53E-13	0.3025	*CLEC1A*GPR142*
rs2306894_BB, rs4078431_BB	-0.8244	4.55E-13	0.0574	*CLEC1A*PRKAG2-AS1*
rs7221462_BB, rs11721292_AA	1.3215	3.09E-11	0.8905	*IFT20*RPL32*
rs2242373_AA, rs4078431_BB	-1.2188	3.51E-13	0.0756	*MFSD6L*PRKAG2-AS1*

544 '*' sign between two gene symbols indicates cross-gene epistasis.

545

Table 7. The comparison of different algorithms.

Algorithm	# Input SNP	Time Cost	Top 15	Top 30	Top 45	Top 60
Classification – Classifying AD patients						
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
Regression – Predicting the progression of AD						
GenEpi	5,086,639	28.38	0.35	0.48	0.46	0.50
BOOST	5,086,639	327.12	0.08	0.08	0.05	0.04
ReliefF	69,842	1.05	0.16	0.22	0.23	0.26

Time 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 in classifying AD patients and predicting the progression of AD are the F1 scores and the average of Pearson and Spearman correlation, respectively.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been supported by the Ministry of Science and Technology of Taiwan grant 105-2911-I-002-566, 105-2221-E-002-129-MY3 and 103-2627-M002-015.

Authors' contributions

Y.-C.C. initiated the study, designed the analysis procedures, performed the analysis and wrote the manuscript. J.-T.W. finished the literature survey of the epistasis that selected by GenEpi and wrote the discussion. C.-Y.C. is the main advisor of Y.-C.C., guided this research on right way and provided ideas to optimize the methods. Y.-C.C, M.-Y.H., Y.-A.T, C.-Y.C. and Y.-J.O. were team members of Alzheimer's disease Dream Challenge. C.-Y.C., Y.-J.O., P.-H.H. K.M.G. and S.W.Y. commented on the draft and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; AmorfixLife Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.;

584 Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company;
585 F.Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare;
586 Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.;
587 Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck &
588 Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.;
589 Servier; Synarc Inc. and Takeda Pharmaceutical Company. The Canadian Institutes of Health
590 Research is providing funds to support ADNI clinical sites in Canada. Private sector
591 contributions are facilitated by the Foundation for the National Institutes of Health
592 (www.fnih.org). The grantee organization is the Northern California Institute for Research and
593 Education and the study is coordinated by the Alzheimer's Disease Cooperative Study at the
594 University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro
595 Imaging at the University of California, Los Angeles. This research was also supported by NIH
596 grants P30 AG010129, K01 AG030514 and the Dana Foundation.

597 ## Reference

598 1. Kingsmore SF, Lindquist IE, Mudge J, Gessler DD, Beavis WD: **Genome-wide**
599 **association studies: progress and potential for drug discovery and development**.
600 *Nature Reviews Drug Discovery* 2008, **7**(3):221-230.
601 2. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M,
602 Nakamura Y *et al*: **Functional SNPs in the lymphotoxin-alpha gene that are**
603 **associated with susceptibility to myocardial infarction**. *Nat Genet* 2002, **32**(4):650-
604 654.
605 3. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni
606 JP, Mane SM, Mayne ST *et al*: **Complement factor H polymorphism in age-related**
607 **macular degeneration**. *Science* 2005, **308**(5720):385-389.
608 4. Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E,
609 Garcia-Garcia J, Sanz F, Furlong LI: **DisGeNET: a comprehensive platform**
610 **integrating information on human disease-associated genes and variants**. *Nucleic*
611 *Acids Res* 2017, **45**(D1):D833-D839.
612 5. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP,
613 Hirschhorn JN: **Genome-wide association studies for complex traits: consensus,**
614 **uncertainty and challenges**. *Nat Rev Genet* 2008, **9**(5):356-369.
615 6. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH: **Missing**
616 **heritability and strategies for finding the underlying causes of complex disease**. *Nat*
617 *Rev Genet* 2010, **11**(6):446-450.
618 7. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy
619 MI, Ramos EM, Cardon LR, Chakravarti A *et al*: **Finding the missing heritability of**
620 **complex diseases**. *Nature* 2009, **461**(7265):747-753.
621 8. Frazer KA, Murray SS, Schork NJ, Topol EJ: **Human genetic variation and its**
622 **contribution to complex traits**. *Nat Rev Genet* 2009, **10**(4):241-251.
623 9. Carlborg O, Haley CS: **Epistasis: too often neglected in complex trait studies?** *Nat*
624 *Rev Genet* 2004, **5**(8):618-625.
625 10. Cordell HJ: **Detecting gene-gene interactions that underlie human diseases**. *Nat Rev*
626 *Genet* 2009, **10**(6):392-404.
627 11. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG,
628 Struewing JP, Morrison J, Field H, Luben R *et al*: **Genome-wide association study**
629 **identifies novel breast cancer susceptibility loci**. *Nature* 2007, **447**(7148):1087-1093.

630 12. Shriner D, Vaughan LK, Padilla MA, Tiwari HK: **Problems with genome-wide**
631 **association studies**. *Science* 2007, **316**(5833):1840-1842.
632 13. Bodmer W, Bonilla C: **Common and rare variants in multifactorial susceptibility to**
633 **common diseases**. *Nat Genet* 2008, **40**(6):695-701.
634 14. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio
635 TA: **Potential etiologic and functional implications of genome-wide association loci**
636 **for human diseases and traits**. *Proc Natl Acad Sci U S A* 2009, **106**(23):9362-9367.
637 15. Moore JH, Asselbergs FW, Williams SM: **Bioinformatics challenges for genome-wide**
638 **association studies**. *Bioinformatics* 2010, **26**(4):445-455.
639 16. Marchini J, Donnelly P, Cardon LR: **Genome-wide strategies for detecting multiple**
640 **loci that influence complex diseases**. *Nat Genet* 2005, **37**(4):413-417.
641 17. Wei WH, Hemani G, Haley CS: **Detecting epistasis in human complex traits**. *Nat Rev*
642 *Genet* 2014, **15**(11):722-733.
643 18. Schupbach T, Xenarios I, Bergmann S, Kapur K: **FastEpistasis: a high performance**
644 **computing solution for quantitative trait epistasis**. *Bioinformatics* 2010,
645 **26**(11):1468-1469.
646 19. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W: **BOOST: A fast approach to**
647 **detecting gene-gene interactions in genome-wide case-control studies**. *Am J Hum*
648 *Genet* 2010, **87**(3):325-340.
649 20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar
650 P, de Bakker PI, Daly MJ *et al*: **PLINK: a tool set for whole-genome association and**
651 **population-based linkage analyses**. *Am J Hum Genet* 2007, **81**(3):559-575.
652 21. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation**
653 **PLINK: rising to the challenge of larger and richer datasets**. *Gigascience* 2015, **4**:7.
654 22. Moore JH, Williams SM: **New strategies for identifying gene-gene interactions in**
655 **hypertension**. *Ann Med* 2002, **34**(2):88-95.
656 23. Yang P, Ho JW, Yang YH, Zhou BB: **Gene-gene interaction filtering with ensemble**
657 **of filters**. *BMC Bioinformatics* 2011, **12 Suppl 1**:S10.
658 24. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P:
659 **Identifying SNPs predictive of phenotype using random forests**. *Genet Epidemiol*
660 2005, **28**(2):171-182.
661 25. Schwarz DF, Konig IR, Ziegler A: **On safari to Random Jungle: a fast**
662 **implementation of Random Forests for high-dimensional data**. *Bioinformatics* 2010,
663 **26**(14):1752-1758.

664 26. Wan X, Yang C, Yang Q, Xue H, Tang NL, Yu W: **MegaSNPHunter: a learning**
665 **approach to detect disease predisposition SNPs and high level interactions in**
666 **genome wide association study**. *BMC Bioinformatics* 2009, **10**:13.
667 27. Murk W, DeWan AT: **Exhaustive Genome-Wide Search for SNP-SNP Interactions**
668 **Across 10 Human Diseases**. *G3 (Bethesda)* 2016, **6**(7):2043-2050.
669 28. Allen GI, Amoroso N, Anghel C, Balagurusamy V, Bare CJ, Beaton D, Bellotti R,
670 Bennett DA, Boehme KL, Boutros PC *et al*: **Crowdsourced estimation of cognitive**
671 **decline and resilience in Alzheimer's disease**. *Alzheimers Dement* 2016, **12**(6):645-
672 653.
673 29. Ma L, Clark AG, Keinan A: **Gene-based testing of interactions in association studies**
674 **of quantitative traits**. *PLoS Genet* 2013, **9**(2):e1003321.
675 30. Oh S, Lee J, Kwon MS, Weir B, Ha K, Park T: **A novel method to identify high order**
676 **gene-gene interactions in genome-wide association studies: gene-based MDR**. *BMC*
677 *Bioinformatics* 2012, **13 Suppl 9**:S5.
678 31. Li S, Cui Y: **Gene-centric gene–gene interaction: A model-based kernel machine**
679 **method**. *The Annals of Applied Statistics* 2012, **6**(3):1134-1161.
680 32. Wu X, Dong H, Luo L, Zhu Y, Peng G, Reveille JD, Xiong M: **A novel statistic for**
681 **genome-wide interaction analysis**. *PLoS Genet* 2010, **6**(9):e1001131.
682 33. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The**
683 **human genome browser at UCSC**. *Genome Res* 2002, **12**(6):996-1006.
684 34. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer
685 TR, Fujita PA, Guruvadoo L, Haeussler M *et al*: **The UCSC Genome Browser**
686 **database: 2015 update**. *Nucleic Acids Res* 2015, **43**(Database issue):D670-681.
687 35. Lewontin RC: **The Interaction of Selection and Linkage. I. General Considerations;**
688 **Heterotic Models**. *Genetics* 1964, **49**(1):49-67.
689 36. Meinshausen N, Bühlmann P: **Stability selection**. *Journal of the Royal Statistical*
690 *Society: Series B (Statistical Methodology)* 2010, **72**(4):417-473.
691 37. Urbanowicz RJ, Meeker M, LaCava W, Olson RS, Moore JH: **Relief-Based Feature**
692 **Selection: Introduction and Review**. *arXiv preprint arXiv:171108421* 2017.
693 38. Urbanowicz RJ, Kiralis J, Sinnott-Armstrong NA, Heberling T, Fisher JM, Moore JH:
694 **GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models**
695 **with random architectures**. *BioData Min* 2012, **5**(1):16.
696 39. Consortium GT: **The Genotype-Tissue Expression (GTEx) project**. *Nat Genet* 2013,
697 **45**(6):580-585.

698 40. Liddelow SA, Guttenplan KA, Clarke LE, Bennett FC, Bohlen CJ, Schirmer L, Bennett
699 ML, Munch AE, Chung WS, Peterson TC *et al*: **Neurotoxic reactive astrocytes are**
700 **induced by activated microglia**. *Nature* 2017, **541**(7638):481-487.
701 41. von Bernhardi R, Ramirez G: **Microglia-astrocyte interaction in Alzheimer's disease:**
702 **friends or foes for the nervous system?** *Biol Res* 2001, **34**(2):123-128.
703 42. Raj D, Yin Z, Breur M, Doorduin J, Holtman IR, Olah M, Mantingh-Otter IJ, Van Dam
704 D, De Deyn PP, den Dunnen W *et al*: **Increased White Matter Inflammation in**
705 **Aging- and Alzheimer's Disease Brain**. *Front Mol Neurosci* 2017, **10**:206.
706 43. Ward LD, Kellis M: **HaploReg: a resource for exploring chromatin states,**
707 **conservation, and regulatory motif alterations within sets of genetically linked**
708 **variants**. *Nucleic Acids Res* 2012, **40**(Database issue):D930-934.
709 44. Li P, Lahvic JL, Binder V, Pugach EK, Riley EB, Tamplin OJ, Panigrahy D, Bowman
710 TV, Barrett FG, Heffner GC *et al*: **Epoxyeicosatrienoic acids enhance embryonic**
711 **haematopoiesis and adult marrow engraftment**. *Nature* 2015, **523**(7561):468-471.
712 45. Kim E, Lee SH, Lee KS, Cheong HK, Namkoong K, Hong CH, Oh BH: **AMPK**
713 **gamma2 subunit gene PRKAG2 polymorphism associated with cognitive**
714 **impairment as well as diabetes in old age**. *Psychoneuroendocrinology* 2012,
715 **37**(3):358-365.
716 46. Peixoto CA, Oliveira WH, Araujo S, Nunes AKS: **AMPK activation: Role in the**
717 **signaling pathways of neuroinflammation and neurodegeneration**. *Exp Neurol* 2017,
718 **298**(Pt A):31-41.
719 47. Whitfield JF, Chiarini A, Dal Pra I, Armato U, Chakravarthy B: **The Possible Roles of**
720 **the Dentate Granule Cell's Leptin and Other Ciliary Receptors in Alzheimer's**
721 **Neuropathology**. *Cells* 2015, **4**(3):253-274.
722 48. Vivier E, Tomasello E, Paul P: **Lymphocyte activation via NKG2D: towards a new**
723 **paradigm in immune recognition?** *Curr Opin Immunol* 2002, **14**(3):306-311.
724 49. Collins RW: **Human MHC class I chain related (MIC) genes: their biological**
725 **function and relevance to disease and transplantation**. *Eur J Immunogenet* 2004,
726 **31**(3):105-114.
727 50. Quiroga I, Lehmann DJ, Barnardo MC, Fuggle S, Cortina-Borja M, Warden DR, Smith
728 AD: **Association study of MICA and MICB in Alzheimer's disease**. *Tissue Antigens*
729 2009, **74**(3):241-243.
730 51. Mirza Z, Rajeh N: **Identification Of Electrophysiological Changes In Alzheimer's**
731 **Disease: A Microarray Based Transcriptomics And Molecular Pathway Analysis**

732 **Study**. *CNS Neurol Disord Drug Targets* 2017.
733 52. Saura CA, Parra-Damas A, Enriquez-Barreto L: **Gene expression parallels synaptic**
734 **excitability and plasticity changes in Alzheimer's disease**. *Front Cell Neurosci* 2015,
735 **9**:318.
736 53. Daschil N, Obermair GJ, Flucher BE, Stefanova N, Hutter-Paier B, Windisch M,
737 Humpel C, Marksteiner J: **CaV1.2 calcium channel expression in reactive astrocytes**
738 **is associated with the formation of amyloid-beta plaques in an Alzheimer's disease**
739 **mouse model**. *J Alzheimers Dis* 2013, **37**(2):439-451.
740 54. Porcellini E, Carbone I, Ianni M, Licastro F: **Alzheimer's disease gene signature says:**
741 **beware of brain viral infections**. *Immun Ageing* 2010, **7**:16.
742 55. Talkowski ME, Maussion G, Crapper L, Rosenfeld JA, Blumenthal I, Hanscom C,
743 Chiang C, Lindgren A, Pereira S, Ruderfer D *et al*: **Disruption of a large intergenic**
744 **noncoding RNA in subjects with neurodevelopmental disabilities**. *Am J Hum Genet*
745 2012, **91**(6):1128-1134.
746 56. Uhrig M, Ittrich C, Wiedmann V, Knyazev Y, Weninger A, Riemenschneider M,
747 Hartmann T: **New Alzheimer amyloid beta responsive genes identified in human**
748 **neuroblastoma cells by hierarchical clustering**. *PLoS One* 2009, **4**(8):e6779.
749 57. Wang W, Mandel J, Bouaziz J, Commenges D, Nabirotchkine S, Chumakov I, Cohen
750 D, Guedj M, Alzheimer's Disease Neuroimaging I: **A Multi-Marker Genetic**
751 **Association Test Based on the Rasch Model Applied to Alzheimer's Disease**. *PLoS*
752 *One* 2015, **10**(9):e0138223.
753 58. Babic Leko M, Borovecki F, Dejanovic N, Hof PR, Simic G: **Predictive Value of**
754 **Cerebrospinal Fluid Visinin-Like Protein-1 Levels for Alzheimer's Disease Early**
755 **Detection and Differential Diagnosis in Patients with Mild Cognitive Impairment**.
756 *J Alzheimers Dis* 2016, **50**(3):765-778.
757 59. Kirkwood CM, MacDonald ML, Schempf TA, Vatsavayi AV, Ikonomovic MD, Koppel
758 JL, Ding Y, Sun M, Kofler JK, Lopez OL *et al*: **Altered Levels of Visinin-Like Protein**
759 **1 Correspond to Regional Neuronal Loss in Alzheimer Disease and**
760 **Frontotemporal Lobar Degeneration**. *J Neuropathol Exp Neurol* 2016, **75**(2):175-182.
761 60. Luo X, Hou L, Shi H, Zhong X, Zhang Y, Zheng D, Tan Y, Hu G, Mu N, Chan J *et al*:
762 **CSF levels of the neuronal injury biomarker visinin-like protein-1 in Alzheimer's**
763 **disease and dementia with Lewy bodies**. *J Neurochem* 2013, **127**(5):681-690.
764 61. Pietrzak M, Rempala G, Nelson PT, Zheng JJ, Hetman M: **Epigenetic silencing of**
765 **nucleolar rRNA genes in Alzheimer's disease**. *PLoS One* 2011, **6**(7):e22585.

766 62. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG,
767 Ma'ayan A: **The harmonizome: a collection of processed datasets gathered to serve**
768 **and mine knowledge about genes and proteins**. *Database (Oxford)* 2016, **2016**.

769

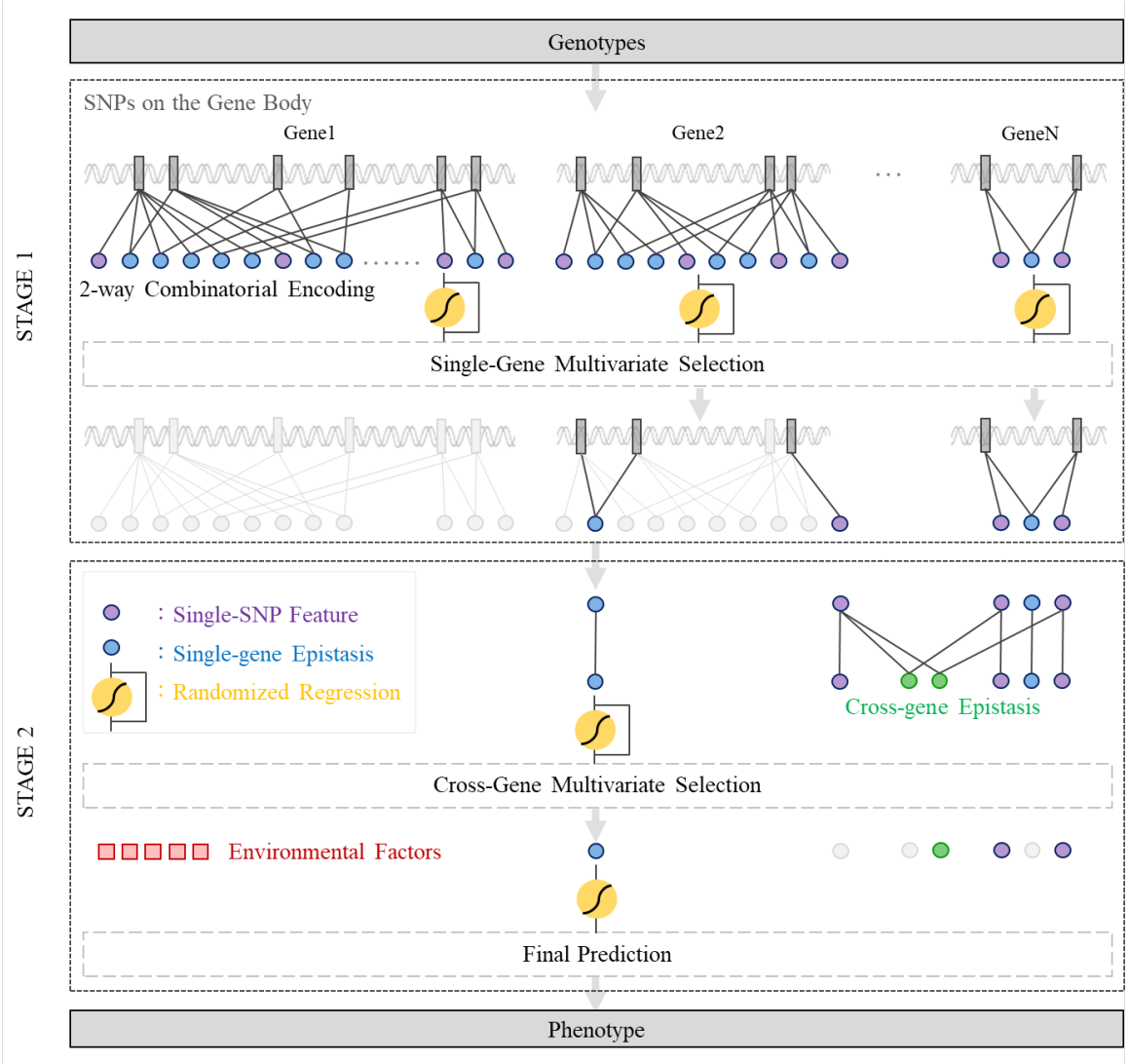

Fig. 1. The architecture of GenEpi.

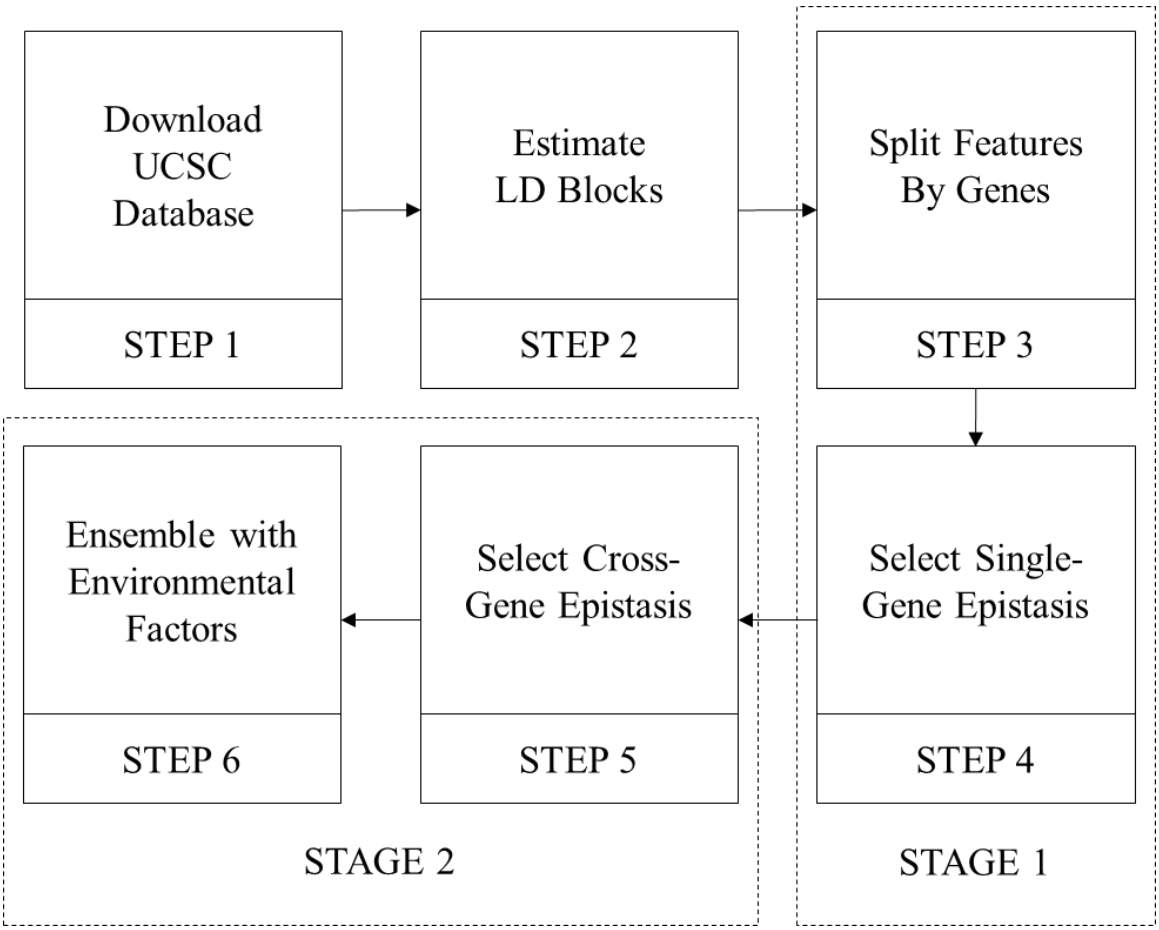

Fig. 2. The modules and data flow in the proposed GenEpi Python package. The first and second modules (STEP 1 and STEP 2) are the two pre-processing steps for downloading the UCSC database and reducing the dimension using LD. The third and fourth modules (STEP 3 and STEP 4) correspond to the STAGE 1 of GenEpi, and the final two modules (STEP 5 and STEP 6) correspond to the STAGE 2 of GenEpi.

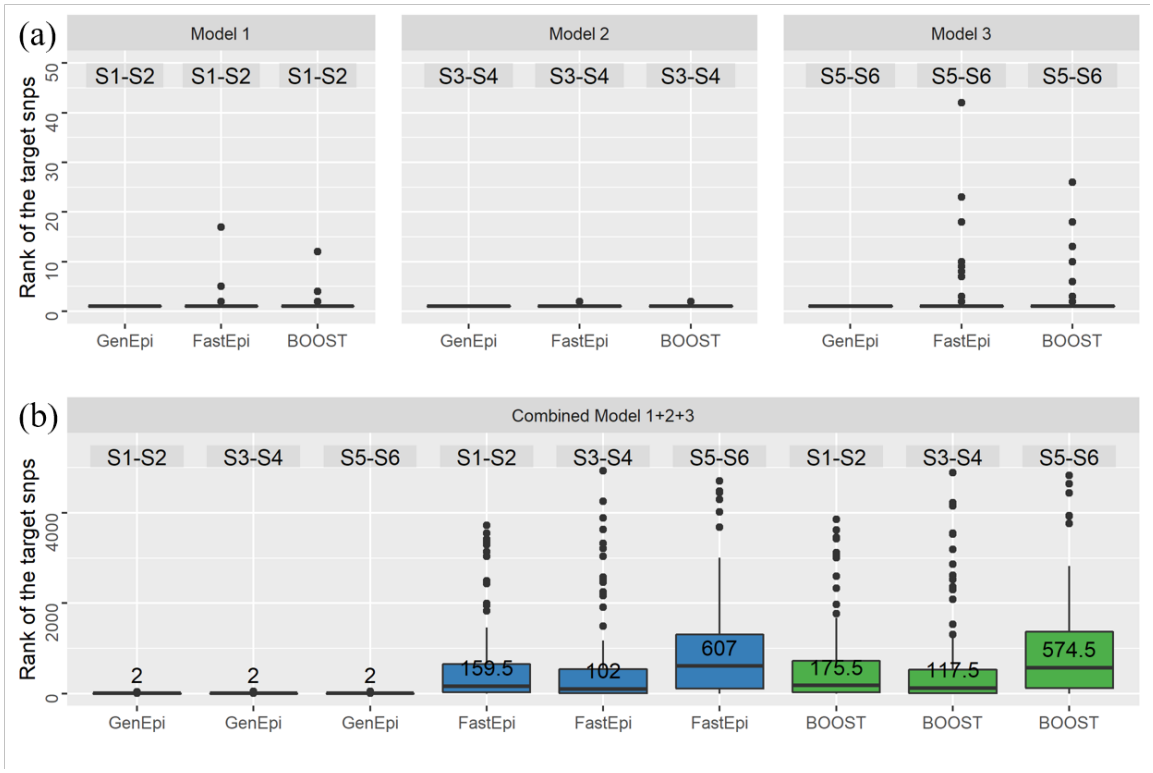

Fig. 3. The boxplot for the ranking of the target epistasis in different algorithms. (a) The upper part are the results of three basic-model datasets with one epistasis consisting of a SNP pair. (b) The lower part is the result of the complex-model dataset, which contained three epistasis. The 'S1-S2' means the epistasis between SNP 1 and SNP 2 and so on. The values on the boxplot are the medians of the ranking of the target epistasis.

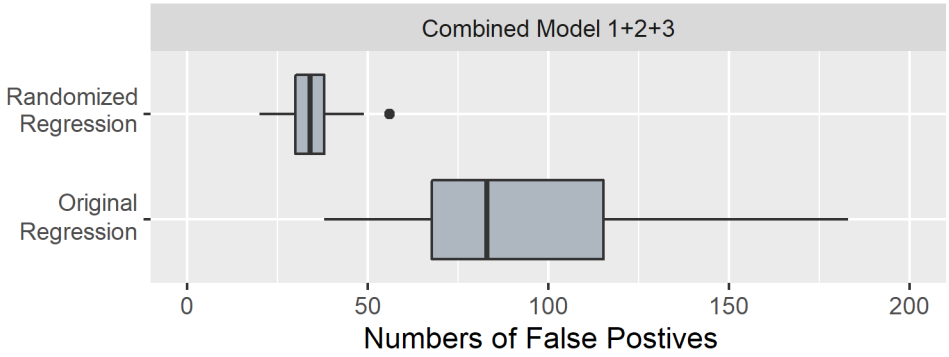

Fig. 4. The numbers of false positives in randomized and original regression.

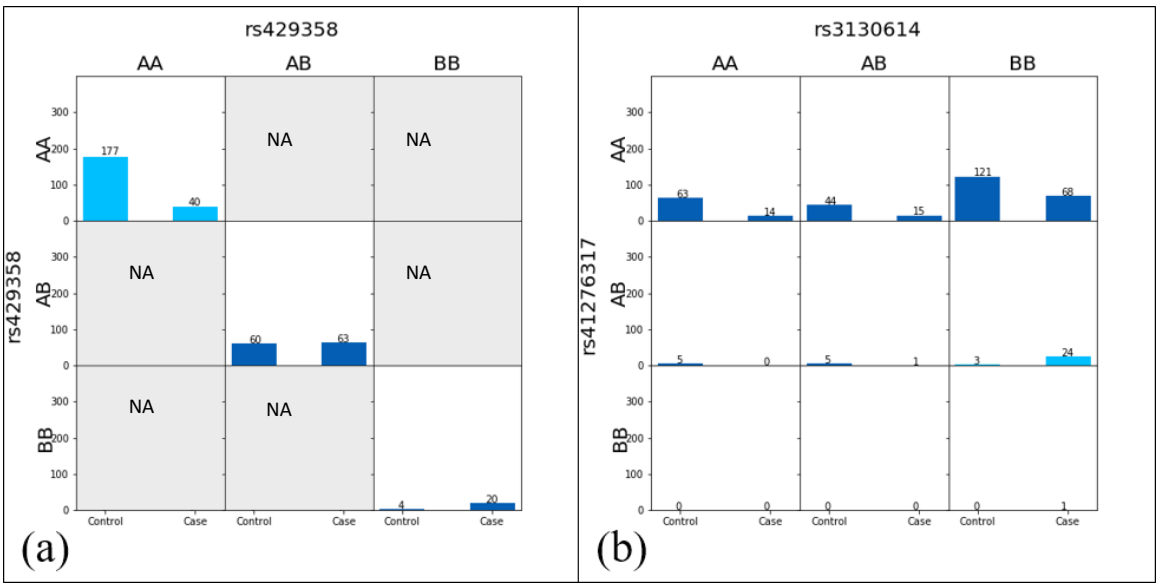

Fig. 5. (a) The bar plot for the population of three allele types in rs429358. (b) The bar plot for combinations of SNPs from the cross-gene epistasis feature selected by GenEpi in predicting control subjects or AD patients. The color 'light blue' is used to highlight the features selected by GenEpi.

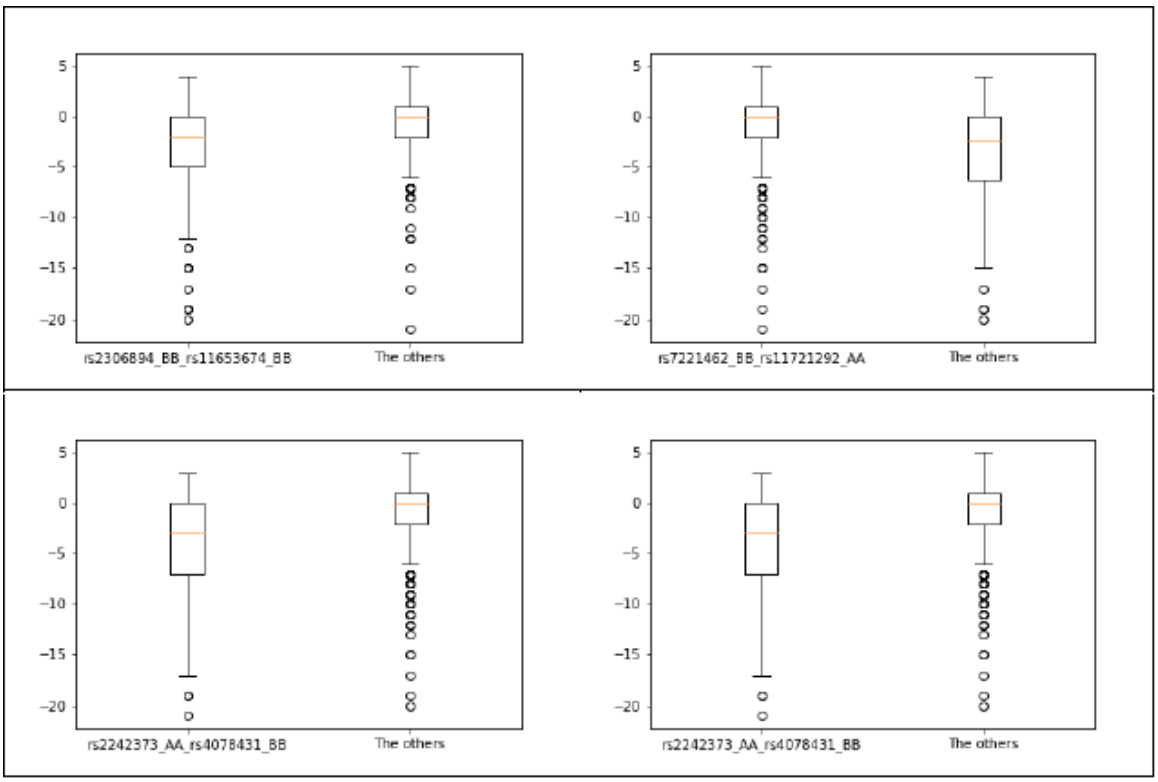

Fig. 6. The box plots for combinations of SNPs from the genetic features selected by GenEpi in predicting the progression of AD. The y-axes of the plot are MMSE difference between '24-month follow-up MMSE' and 'MMSE at baseline'.

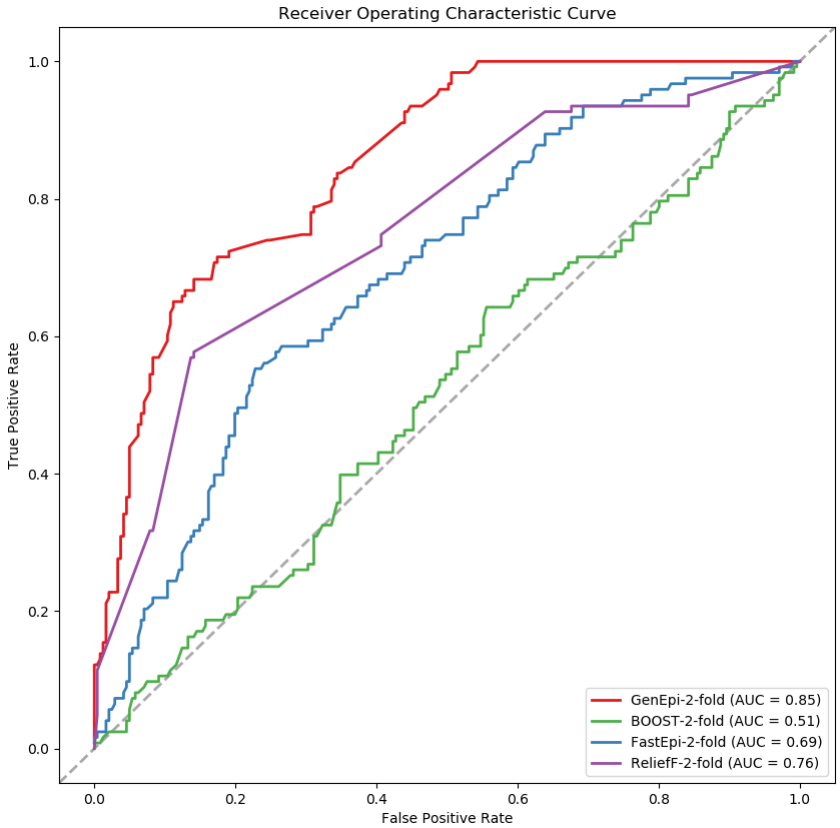

Fig. 7. The ROC curves of different algorithms.

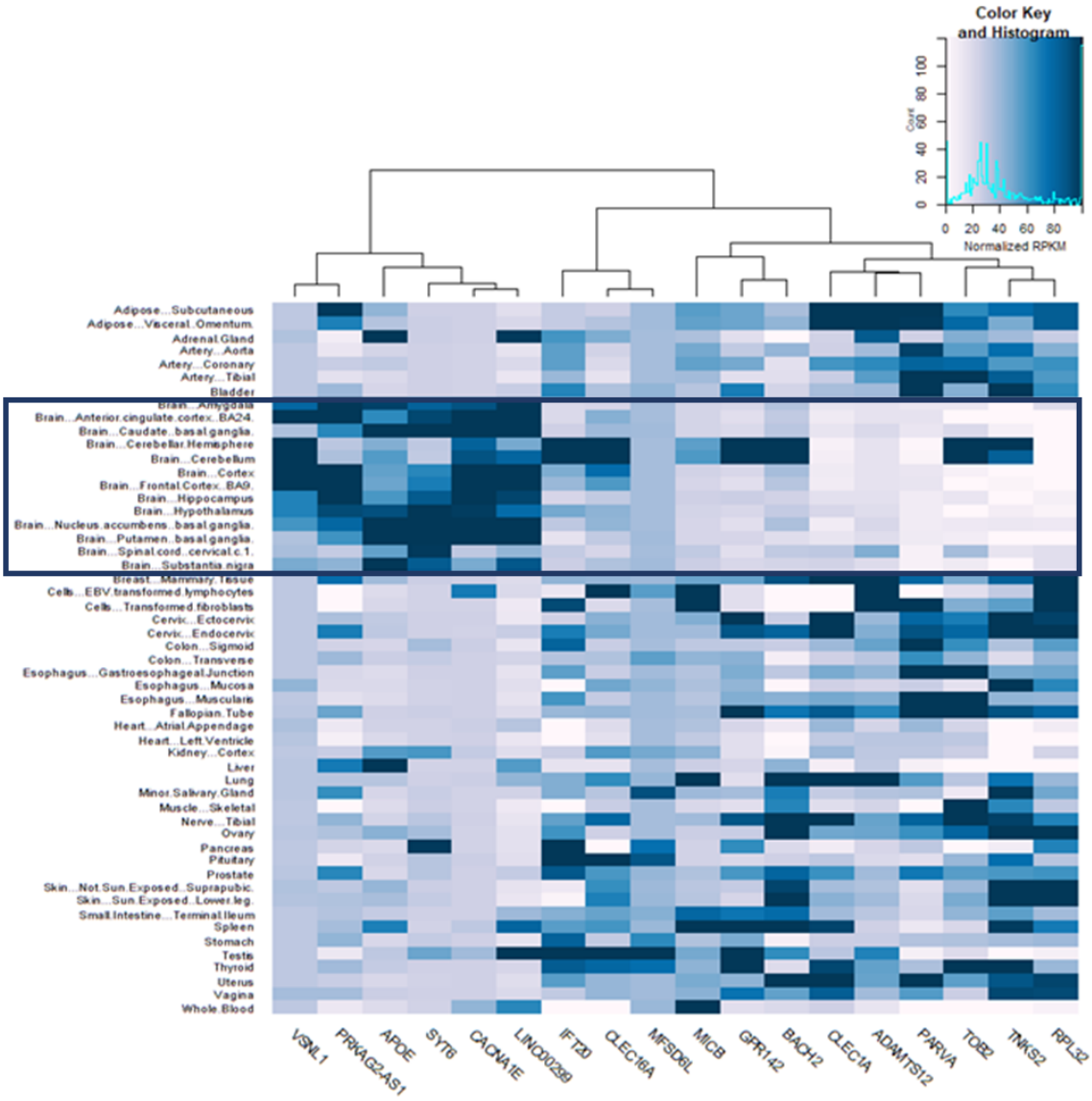

Fig. 8. The heatmap of gene expression in different tissues for the 18 genes selected by GenEpi. The blue box highlights the sub-regions of brain.