BMC Genetics

RESEARCH ARTICLE Open Access

MCC-SP: a powerful integration method for identification of causal pathways from genetic variants to complex disease

Yuchen Zhu[1†], Jiadong Ji[2†], Weiqiang Lin[1], Mingzhuo Li[1], Lu Liu[1], Huanhuan Zhu[3,4], Fuzhong Xue[1], Xiujun Li[1], Xiang Zhou[3,4] and Zhongshang Yuan[1*]

Abstract

Background: Genome-wide association studies (GWAS) have successfully identified genetic susceptible variants for complex diseases. However, the underlying mechanism of such association remains largely unknown. Most disease-associated genetic variants have been shown to reside in noncoding regions, leading to the hypothesis that regulation of gene expression may be the primary biological mechanism. Current methods to characterize gene expression mediating the effect of genetic variant on diseases, often analyzed one gene at a time and ignored the network structure. The impact of genetic variant can propagate to other genes along the links in the network, then to the final disease. There could be multiple pathways from the genetic variant to the final disease, with each having the chain structure since the first node is one specific SNP (Single Nucleotide Polymorphism) variant and the end is disease outcome. One key but inadequately addressed question is how to measure the between-node connection strength and rank the effects of such chain-type pathways, which can provide statistical evidence to give the priority of some pathways for potential drug development in a cost-effective manner.

Results: We first introduce the maximal correlation coefficient (MCC) to represent the between-node connection, and then integrate MCC with K shortest paths algorithm to rank and identify the potential pathways from genetic variant to disease. The pathway importance score (PIS) was further provided to quantify the importance of each pathway. We termed this method as "MCC-SP". Various simulations are conducted to illustrate MCC is a better measurement of the between-node connection strength than other quantities including Pearson correlation, Spearman correlation, distance correlation, mutual information, and maximal information coefficient. Finally, we applied MCC-SP to analyze one real dataset from the Religious Orders Study and the Memory and Aging Project, and successfully detected 2 typical pathways from APOE genotype to Alzheimer's disease (AD) through gene expression enriched in Alzheimer's disease pathway.

Conclusions: MCC-SP has powerful and robust performance in identifying the pathway(s) from the genetic variant to the disease. The source code of MCC-SP is freely available at GitHub (https://github.com/zhuyuchen95/ADnet).

Keywords: Maximum correlation coefficient, K shortest paths algorithms, Integration method, Pathway, Alzheimer's disease

* Correspondence: yuanzhongshang@sdu.edu.cn
†Yuchen Zhu and Jiadong Ji contributed equally to this work.
[1]Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, Shandong, China
Full list of author information is available at the end of the article

Background

Over the last decade, genome-wide association studies (GWAS) have achieved remarkable successes in identifying genetic susceptible variants (e.g. SNPs, Single Nucleotide Polymorphism) for a variety of complex traits or diseases [1]. However, the underlying biological pathway mechanism of such association remains largely unknown. Indeed, the genetic variant can act on other molecular traits (e.g. gene expression) and they together weave into one biological network or pathway that contributes to a disease. For instance, the *APOE* gene has been well identified to be associated with Alzheimer's disease from large scale GWAS [2–5], one possible explanation is that the SNP variant in *APOE* can first regulate the *APOE* gene expression [6–10], then act on the production of amyloid plaques and neurofibrillary tangles, and finally lead to AD [11–13].

Most disease-associated genetic variants from GWAS have been shown to lie in noncoding regions across the genome [1, 14, 15], which provides the clues that regulation of gene expression levels may be the primary biological mechanism through which genetic variants affect complex disease. Certainly, the top GWAS SNP can be also significantly detected due to the linkage disequilibrium with the true causal one, which could be an exonic variant [16]. In addition, several expression quantitative trait loci (eQTLs) studies also illustrate that the expression regulatory information may play a pivotal role in bridging the gap between genetic variants and traits [17–19]. Up until now, there are a few methods to characterize the gene expression that mediates the effect of the genetic variant on complex disease. Huang et al. proposed a model to exploit gene expression to more powerfully test the association between SNPs and diseases by jointly modeling SNPs, gene expressions and diseases [20, 21]. Recent transcriptome-wide association studies (TWAS) have been widely used to integrate the expression regulatory information from eQTL studies with GWAS data to identify gene expression that links the cis-SNPs (SNPs that are within a predefined gene or other well-defined genetic region) and the complex disease [22–24]. Nevertheless, these studies commonly analyzed one gene at a time, while the genetic variant can affect the complex disease through multiple genes and multiple pathways. Park et al. developed the causal multivariate mediation within extended linkage disequilibrium (CaMMEL) method in Bayesian inference framework to select target genes mediating the effect of genetic variants on the complex disease [25]. Wei and Li proposed the nonparametric pathways-based regression (NPR) that can consider multiple pathways simultaneously and allow complex interactions among genes within the pathways [26]. Yao et al. developed a model to quantify the proportion of disease heritability that is specifically mediated in *cis* region by the assayed expression levels of the set of all genes, and of genes in specific functional categories [27]. Indeed, it has been well documented in GWAS that the multiple gene or pathway-based approach can improve power [28]. Although these methods include multiple gene expression in the model, they ignore the complex network structure relationship among genes, which, from the network medicine perspective, is hard to investigate the precise network or molecular pathways involved in complex disease. In fact, one single gene expression can express some mediated effects from the SNP variant to the disease when studying it alone, while this effect could change substantially when studying it within one network or pathway, and vice versa [29, 30]. The focus has been shifted to the identification of pathways.

Often, there could be multiple pathways from the SNP variant to the final disease, with each having the chain structure since the first node is specific SNP variant and the end outcome is the disease. One key but inadequately addressed question is, given such chain pathway, how to measure the connection strength between two nodes and detect whether such pathway is the potential causal one. It is highly desirable to develop statistical methods for ranking the effect of these pathways, providing evidence to give the priority of some pathways for potential drug development and offer drug targets in a cost-effective and timely manner. In the context of systems epidemiology, Ji et al. developed a statistic to test the pathway effect that contributed to a disease with a case-control design [31]. Yuan et al. proposed a novel chi-square statistic to identify whether one chain-type pathway is associated with the final disease [30]. However, their methods simply use the linear regression to represent the between-node correlation, which is insufficient to capture the complex dependency between the nodes. Furthermore, their methods did not include the outcome variable (complex trait or disease) into the pathway and cannot essentially investigate the pathway mechanism. For example, if there is one potential pathway SNP → gene expression 1 → gene expression 2 → AD, their methods, under linear between-node correlation, can detect SNP → gene expression 1 → gene expression 2 is significantly associated with AD, but failed to determine whether AD is connected to gene expression 1 or gene expression 2. Given that the goal is to rank and quantify the effect of the pathways from the genetic variant to the complex disease, it is intuitively to put such a question into the framework of graph theory once the suitable quantity is found to measure the connection strength between two nodes in the pathway. At present study, we first introduced the maximal correlation coefficient (MCC) to represent the between-node connection, and then integrate MCC with the commonly used K shortest paths algorithm [32–34] in graph theory to rank and identify the potential pathways from genetic variant to disease. We further defined the pathway importance score (PIS) to quantify the importance of each pathway. We termed this method as "MCC-SP". Various simulations, with different sample sizes and network structures, are conducted to illustrate MCC is better to measure the between-node correlation than other quantities including Pearson correlation, Spearman correlation, distance correlation, mutual information, and maximal information coefficient. MCC-SP, as an integration method, has always better and robust performance in identifying the causal pathway from genetic variant to the disease. From the Religious Orders Study and the Memory and Aging Project (ROSMAP), we further applied MCC-SP to identify the potential causal pathway from *APOE* genotype to AD through gene expression enriched in Alzheimer's disease pathway.

Results

Simulation

Table 1 shows the simulation results when all the between-node correlations are linear. When the sample size is relatively large (e.g. 300, 500), all methods except MI-SP and MIC-SP have comparable performance as the Pearson-SP, which is the gold standard in such case under both all-right and range-right criteria. When the sample size reduced to 100, the superiority of Pearson -SP is more obvious, though the power of all methods decreases. Figures 1 and 2 show the results of all six integration-methods under sample size 500 and 4 different nonlinear correlation patterns being arcuate, cosine, quadratic and mixed pattern. Under the arcuate nonlinear pattern, the MCC-SP performs better than any other method under both criteria regardless of the proportion of nonlinear components (Figs. 1a and 2a). Note that under the all-right criteria, the other methods are unable to identify the top 4 pathways at all. Under cosine nonlinear relationship, both MCC-SP and DC-SP have comparably better performance than that of the other methods (Figs. 1b and 2b), even when the proportion of nonlinear component reached 60% (Figure S1). However, MCC-SP has the best performance under the all-right criteria when the nonlinear proportion is 30% (Figure S2). Similar phenomenon can be found under the sine relationship (Figure S3). Under the mixed nonlinear relationship, MCC-SP performs best under both criteria (Fig. 1c) and have comparable better performance with DC-SP than that of other methods (Fig. 2c). Under the quadratic relationship, MCC-SP still have the best performance than any other methods (Fig. 1d and Fig. 2d). The results are consistent under the exponential relationship or the reciprocal relationship (Figure S4), or when comparing MCC-SP with the nonparametric pathways-based regression (NPR) model (Figure S5).

Table 1 The number of times that properly pinpoint the top 4 pathways among 500 simulations with linear between-node correlation

Sample	Criteria	Pearson	Spearman	Distance	MCC	MIC	MI
100	All-right	152	129	125	107	44	1
	Range-right	408	402	428	378	300	46
300	All-right	271	252	248	248	88	0
	Range-right	492	490	495	486	468	218
500	All-right	444	429	415	440	152	254
	Range-right	500	500	500	500	500	500

The simulation results under sample size 100 and 300 are in the Figures S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, where similar phenomenon can be found.

Real data

Overall, totally 6 pathways from *APOE* genotype to AD have been identified to be top 5 from all the 6 integration methods (Table 2). The findings of each method are inconsistent. Pearson-SP, Spearman-SP, DC-SP and MCC-SP ranked the pathway *APOE* genotype →*APOE* gene expression →*GRIN2A* → *CAPN2* → *MAPT*→AD to be first, which illustrates this pathway may play important roles in the AD mechanism. Both DC-SP and MCC-SP ranked *APOE* genotype →*APOE* gene expression →*GRIN2A* → *NOS1* → AD to be second, and this pathway has been ranked to be first by MIC-SP and MI-SP, which indicates that this pathway may also has a high probability to involve the AD mechanism. Both Spearman-SP and MCC-SP ranked *APOE* genotype →*APOE* gene expression →*CACNA1C* → *CAPN2* → *MAPT*→AD as the third. MCC-SP ranked *APOE* genotype →*APOE* gene expression→*CACNA1C* → *NOS1* → AD as the fourth and *APOE* genotype →*APOE* gene expression→*GRIN2A* → *CAPN2* → *MAPT*→AD as the fifth. Under the MCC between-node correlation, the PIS for these top 5 pathways are 13.80, 12.32, 8.87, 8.84 and 7.42 respectively. The top 2 pathways have comparable PIS, which are much higher than that of other pathways. Thus, the top 2 pathways are likely essential for AD development and can be chosen for further experimental verifications. The detailed results for the rank of the total 33 pathways are presented in the Table S1.

Discussion

Most disease-associated genetic variants lie in the noncoding regions of the genome and gene expression levels can bridge the gap between genetic variant and disease. Often, there are multiple pathways from the genetic variant to the final disease, we have developed a powerful integration method, MCC-SP, to rank and identify these multiple potential pathways. Various simulations under

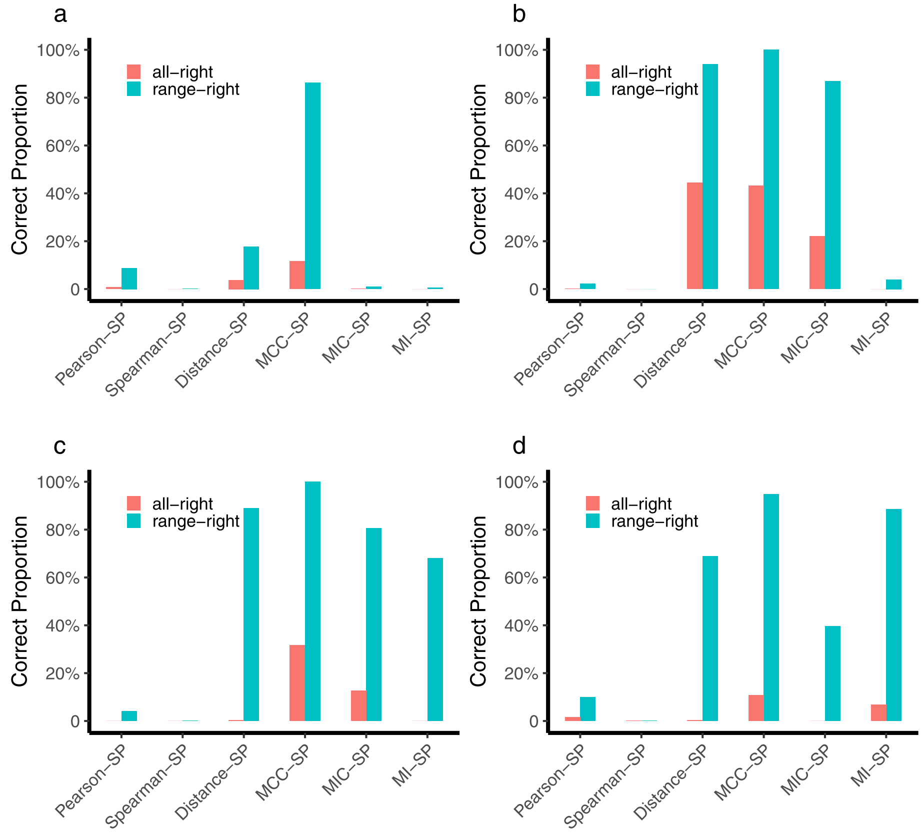

Fig. 1 The proportion that correctly pinpoint the top 4 pathways among 500 simulations under two criteria when the sample size is 500 and the proportion of nonlinear components is 40%.The nonlinear pattern is (**a**) $\boldsymbol{\varphi}(\boldsymbol{x}_i)=\sqrt{\boldsymbol{C}-\boldsymbol{x}_i^2}+\boldsymbol{\varepsilon}$, (**b**) $\boldsymbol{\varphi}(\boldsymbol{x}_i) = \mathbf{cos}\,(\boldsymbol{x}_i)+\boldsymbol{\varepsilon}$, (**c**) mixed nonlinear pattern(6 edges having cosine and 3 edges having quadratic relationship) and (**d**) $\boldsymbol{\varphi}(\boldsymbol{x}_i)=\boldsymbol{x}_i^2+\boldsymbol{\varepsilon}$ respectively

different sample size and different between-node correlation pattern have shown that the proposed method has better performance than other competing methods. ROSMAP data analysis illustrates that the method can partially detect the mechanism from *APOE* genotype to AD through gene expression enriched in AD pathway. The term “integration” here can be interpreted that we have integrated the suitable between-node correlation with the K shortest paths algorithm in graph theory. MCC-SP is essentially network-based and has different model assumptions from traditional TWAS. Statistically, it will lose efficiency if we know the network structure while ignore it during the inference. In this sense, MCC-SP provided an alternative and complement from TWAS to dissect the pathway mediating the genetic variant and the complex disease. MCC-SP are not only limited to gene expression but can be extended to other molecular phenotypes (e.g. proteomics). Note that the relationship among genes may be a mixture of many possible so called “correlations” rather than endorsed to one of the six suggested functions only. Currently, it is hard to extend MCC-SP to summary-level data. For summary level data, one key step is to calculate the correlation matrix among the genes and the traits. If the correlation is linear, we can implement this using some well-known methods [35, 36]. However, as we show here, various nonlinear relationships exist and it is hard to calculate the complex non-linear correlation matrix using summary-level data.

ROSMAP data analysis has found that *APOE* genotype is significant associate with AD ($OR = 2.8849$, $P = < 0.0001$), it is reasonable to utilize gene expression enriched in AD disease pathway to rank and identify the potential pathway from *APOE* to AD. We chose the overlapped genes between the ROSMAP data and those located on the AD disease pathway into the analysis. The top 2 pathways have comparable and much higher pathway important scores (PIS) than other pathways, which indicates that these two pathways may play important roles in AD development. The most important pathway identified is *APOE* genotype →*APOE* gene expression→*GRIN2A* → *CAPN2* → *MAPT*→AD. The *APOE* genotype can regulate its gene expression, in the central nervous system (CNS), *LDLR* family is intimately involved in neuronal signal transduction, modulation of ligand-gated ion channels, and regulating neurite

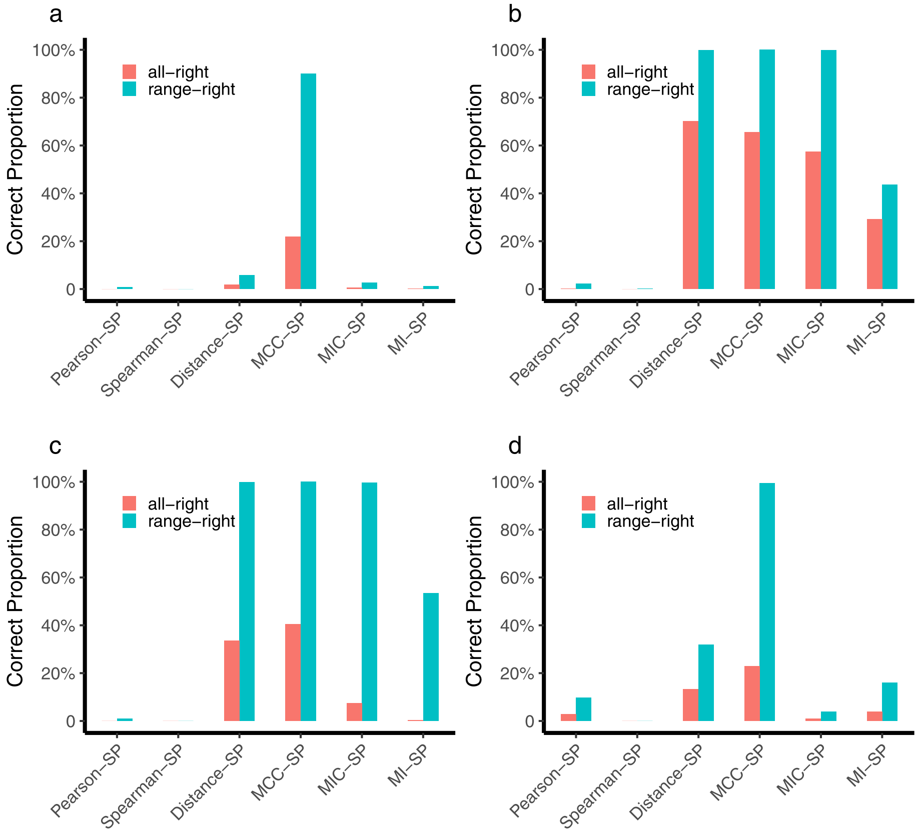

Fig. 2 The proportion that correctly pinpoint the top 4 pathways among 500 simulations under two criteria when the sample size is 500 and the proportion of nonlinear components is 50%.The nonlinear pattern is (**a**) $\boldsymbol{\varphi}(\boldsymbol{x}_i)=\sqrt{\boldsymbol{C}-\boldsymbol{x}_i^2}+\boldsymbol{\varepsilon}$, (**b**) $\boldsymbol{\varphi}(\boldsymbol{x}_i) = \mathbf{cos}\,(\boldsymbol{x}_i)+\boldsymbol{\varepsilon}$, (**c**) mixed nonlinear pattern(8 edges having cosine and 4 edges having quadratic relationship) and (**d**) $\boldsymbol{\varphi}(\boldsymbol{x}_i)=\boldsymbol{x}_i^2+\boldsymbol{\varepsilon}$ respectively

outgrowth, synapse formation and neuronal migration. *ApoE* binds to the highly conserved low-density lipoprotein receptor *(LDLR*) family [37] including *LRP1* and *ApoER2*, while *ApoER2* was reported to bind *NMDAR* (*GRIN2A* belongs to this family) [38, 39]. The *NMDAR* is a cation channel highly permeable to calcium and plays critical roles in governing normal and pathologic functions in neurons. Calcium entry through *NMDAR* can lead to the activation of the Ca2 + −dependent protease, calpain [39, 40]. Gene *MAPT* belongs to the family of *Tau* and *CAPN2* belongs to the family of *Calpain. Calpain*-mediated tau cleavage can play an important role under neurodegenerative conditions [38, 41]. It has been shown that calpain activation results in the generation of several N-terminal tau fragments, which can be detected in mitochondria present in synaptosomal fractions obtained from AD brains [38, 41, 42]. In addition, overexpression of *NMDAR2B* in an inflammatory model of Alzheimer's disease, which can be modulated by *NOS* (*NOS1* belongs to this family) inhibitors [43]. Further independent sample validation and experimental study can be conducted to validate these findings.

One limitation of our method is that we assume the network or pathway structure is assumed to be known (e.g. AD pathway in our ROSMAP data analysis). Little attention has been paid on the network structure learning problem, which means determining every between-node link with highest degree of data matching, and often one joint distribution of variables can reflect more than one network structure. Actually, most biologists and clinical researchers usually have some prior on the interplay between the biological components and can depict more or less the specific network or pathway for the corresponding biological process. Meanwhile, numerous databases (e.g. KEGG) can be further borrowed to establish the network structure. Even so, MCC-SP is unable to deal with the loop network. Another limitation is that there is lack of test for the significance of the pinpointed pathway, for example, the proposed method is unable to test whether the order of identified pathways is significant or not. Some nonparametric techniques (e.g. permutation and bootstrap) may be further developed to solve such problems. In practice, once we have obtained the rank of the pathways, one key question is which pathway should be selected for further

Table 2 The top 5 pathways identified by each method from APOE genotype to AD in ROSMAP study

Method	Order				
	1	2	3	4	5
Pearson-SP	$Path_1$ (152.8)	$Path_2$ (69.90)	$Path_3$ (69.01)	$Path_4$ (59.54)	$Path_5$ (32.35)
Spearman-SP	$Path_1$ (246.54)	$Path_2$ (154.98)	$Path_4$ (125.08)	$Path_5$ (96.00)	$Path_3$ (66.61)
DC-SP	$Path_1$ (57.84)	$Path_3$ (41.32)	$Path_2$ (35.91)	$Path_4$ (31.42)	$Path_6$ (24.32)
MIC-SP	$Path_3$ (3063.17)	$Path_1$ (1136.26)	$Path_6$ (1099.96)	$Path_2$ (743.01)	$Path_4$ (380.01)
MI-SP	$Path_3$ (60.95)	$Path_6$ (39.72)	$Path_2$ (33.01)	$Path_5$ (25.26)	$Path_1$ (21.32)
MCC-SP	$Path_1$ (13.80)	$Path_3$ (12.32)	$Path_4$ (8.87)	$Path_6$ (8.84)	$Path_2$ (7.42)

Note: The pathway importance scores (PISs) were shown in the parenthesis. Note that the PIS is only comparable across one specific method

***Path*$_1$**: *APOE* genotype →*APOE* gene expression→*GRIN2A* → *CAPN2* → *MAPT*→AD

***Path*$_2$**: *APOE* genotype →*APOE* gene expression →*GRIN2A* → *MAPK1* → *CASP3* → AD

***Path*$_3$**: *APOE* genotype →*APOE* gene expression →*GRIN2A* → *NOS1* → AD

***Path*$_4$**: *APOE* genotype →*APOE* gene expression →*CACNA1C* → *CAPN2* → *MAPT*→AD

***Path*$_5$**: *APOE* genotype →*APOE* gene expression→*CACNA1C* → *MAPK1* → *CASP3* → AD

***Path*$_6$**: *APOE* genotype →*APOE* gene expression→*CACNA1C* → *NOS1* → AD

experimental verification. Here we have provided the PIS to quantify the importance of each pathway, we use q_{50} as the threshold to indicate that those pathways with effect greater than the median value, will have the PIS greater than one. Regardless of the threshold, PIS is essentially the product of the "correlation" along the specific pathway. Indeed, even for two pathways with similar PIS, MCC-SP still give the different ranks. For example, in our real data analysis, the PIS for $Path_3$ and $Path_4$ are quite close (8.87 vs 8.84). This indicates that these two pathways may have equivalent effect and importance but with different ranks. In practice, we recommend choosing those pathways having comparable PIS with the top one for further experimental verification in a cost-effective manner.

Conclusions

We proposed an integration method called MCC-SP, identifying the causal pathway effect within a network from genetic variant to the disease. MCC-SP is effective and powerful at identifying the specific pathways contributing that cause disease, and can rank these potential pathways, so it can provide new insights into underlying mechanisms and can provide a more comprehensive approach to studying the effects of specific pathways on disease.

Method

Six between-node correlation measures

Pearson correlation coefficient

The Pearson correlation coefficient is used as an indicator to measure the strength of linear correlation between two random variables X and Y.

$$r=\frac{\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{\sqrt{\sum\left(X-\overline{X}\right)^2\sum\left(Y-\overline{Y}\right)^2}},$$

where $\overline{X},\overline{Y}$ represent the mean of the two variables respectively. The range of r is $[-1,1]$. When $r=0$, there is no linear relationship between the two variables.

Spearman correlation coefficient

The Spearman correlation coefficient indicates the degree of monotonic correlation between two variables X and Y, which is essentially a linear correlation of the ranks of X and Y. It is free of the distribution of variables and is defined as follows:

$$\rho=\frac{\sum_{i=1}^{n}(r_i-\overline{r})(s_i-\overline{s})}{\sqrt{\sum_{i=1}^{n}(r_i-\overline{r})^2}\sqrt{\sum_{i=1}^{n}(s_i-\overline{s})^2}},$$

where r_i (s_i) represents the rank of x_i (y_i) in sample X (Y), and the range of ρ is $[-1,1]$. When $\rho=0$, there is no monotonic relationship between the two variables. When $\rho>0$, the relationship between the two variables increases monotonically; when $\rho<0$, it decreases monotonically.

Distance correlation

The distance-related $R(X,Y)$ is different from the previous correlations based on the covariance matrix and variance matrix. It measures the correlation between variables by calculating the Euclidean distance of the sample itself. $R(X,Y)$ is non-negative and can be used to measure the correlation between X and Y with any dimension. $R(X,Y)=0$ indicates that X and Y are independent.

Suppose the sample data is $(X,Y)=\{(X_k,Y_k):k=1,\ldots,n\}$, and define the following quantities

$$\begin{aligned}a_{kl}&=|X_k-X_l|_p,\ \ \overline{a}_{k\cdot}=\frac{1}{n}\sum\nolimits_{l=1}^{n}a_{kl},\ \ \overline{a}_{\cdot l}\\&=\frac{1}{n}\sum\nolimits_{k=1}^{n}a_{kl},\ \ \overline{a}_{\cdot\cdot}=\frac{1}{n^2}\sum\nolimits_{k,l=1}^{n}a_{kl},A_{kl}\\&=a_{kl}-\overline{a}_{k\cdot}-\overline{a}_{\cdot l}+\overline{a}_{\cdot\cdot}.\end{aligned}$$

$$\begin{aligned}b_{kl}&=|Y_k-Y_l|_q,\ \ B_{kl}=b_{kl}-\overline{b}_{k\cdot}-\overline{b}_{\cdot l}+\overline{b}_{\cdot\cdot},\ \ k,l\\&=1,2,\ldots,n.\end{aligned}$$

The empirical distance covariance is defined as

$$V_n^2(X,Y)=\frac{1}{n^2}\sum\nolimits_{k,l=1}^{n}A_{kl}B_{kl},$$

Similarly, $V_n(X)$ is defined by

$$V_n^2(X)=V_n^2(X,X)=\frac{1}{n^2}\sum\nolimits_{k,l=1}^{n}A_{kl}^2.$$

The empirical distance correlation coefficient $R_n(X, Y)$ is defined as:

$$R_n(X,Y)=\sqrt{R_n^2(X,Y)}=\begin{cases}\sqrt{\dfrac{V_n^2(X,Y)}{\sqrt{V_n^2(X)V_n^2(Y)}}}, & V_n^2(X)V_n^2(Y)>0\\ 0 & V_n^2(X)V_n^2(Y)=0\end{cases}$$

Mutual information based on kernel density estimation (MI)

Mutual Information is a useful measure of information in information theory. It can be seen as the amount of information about a random variable contained in a random variable, or a random variable due to the knowledge of another random variable. Mutual Information does not need to make any assumption about the nature of the relationship between variable characteristics and is defined as

$$I(X,Y)=\int_Y\int_X p(x,y)\ log\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy,$$

where $p(x, y)$ is the joint probability density function of (X, Y), and $p(x)$, $p(y)$ are the corresponding marginal density functions of (X, Y), respectively. $I(X, Y)$ can be viewed as the expected value of $\log(\frac{p(x,y)}{p(x)p(y)})$ (Point Mutual Information, PMI), which is

$$I(X,Y)=E\left(\log\left(\frac{\mathrm{p}(x,y)}{p(x)p(y)}\right)\right).$$

The value of mutual information can also be expressed as Kullback–Leibler divergence (also known as relative entropy),

$$I(X,Y)=H(Y)-H(Y|X),$$

where $H(Y)$ is the entropy of Y, referring to the uncertainty of Y, and $H(Y| X)$ is the uncertainty of Y given X. Thus, $I(X, Y)$ can be interpreted as a quantity introduced by X to reduce the uncertainty of Y. Therefore, the closer the relationship between X and Y, the greater the $I(X, Y)$. $I(X, Y) = 0$ when two variables are independent.

The calculation of MI requires the estimation of the density functions, we adopt the kernel density estimation method. For the one-dimension marginal density, we assume that the data x_1, x_2, …, x_n, are taken from the continuous distribution $p(x)$. A kernel density estimate is defined as

$$\hat{p}(x)=\frac{1}{nh}\sum\nolimits_{i=1}^{n}\omega_i=\frac{1}{nh}\sum\nolimits_{i=1}^{n}K\left(\frac{x-x_i}{h}\right),$$

where h is the bandwidth and $K(\cdot)$ is the kernel function, $K(x) > 0$, $\int K(x)dx = 1$. At present study, we chose the commonly used *Gauss* kernel function.

For the two-dimension joint density, assuming that the data X, Y be a bivariate sample drawn from a common distribution described by the density function. The bivariate kernel density estimation can be defined to be

$$\hat{f}_H(\boldsymbol{z};\boldsymbol{H})=\frac{1}{n}\sum\nolimits_{i=1}^{n}K_H(\boldsymbol{z}-\boldsymbol{Z_i}),$$

where $=(\boldsymbol{x},\boldsymbol{y})^T\ \boldsymbol{Z_i}=(\boldsymbol{X_{i1}},\boldsymbol{Y_{i2}})^T$, $i = 1, 2, …, n$.

Here $\boldsymbol{H}$ is the bandwidth (or smoothing) 2×2 matrix which is symmetric and positive definite; $K(\cdot)$ is the bivariate kernel function which is a symmetric multivariate density and $K_H(\boldsymbol{z})=|\boldsymbol{H}|^{-\frac{1}{2}}K(\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{z})$. At present study, we use the standard multivariate normal kernel: $K_H(\boldsymbol{z})=(2\pi)^{-\frac{d}{2}}|\boldsymbol{H}|^{-\frac{d}{2}}\exp(-\frac{1}{2}\boldsymbol{z}^TH^{-1}\boldsymbol{z})$ with d = 2.

Maximal information coefficient (MIC)

The idea of the MIC is that if there is a relationship between two variables, one can draw a grid on the scatter plot of the two variables, which partitions the data to encapsulate that relationship. To calculate the MIC of a set of two variables, we explore all grids up to a maximal grid resolution which is dependent on the sample size, computing for each pair of integers (X, Y) the largest possible mutual information achievable by any X -by- Y grid applied to the data [44]. These mutual information values are normalized such that the grids of different dimensions are comparable (values between 0 and 1). We define the characteristic matrix M ($M = (m_{X,\ Y})$), where $m_{X,\ Y}$ is the largest normalized mutual information value in all grids, and MIC is the largest value in M. Given the current data set D, the largest mutual information value is recorded as $I(D, X, Y)$, which can be further standardized as

$$\mathrm{M}(D)_{x,y}=\frac{I(D,x,y)}{\log(\ min\{x,y\})},$$

$\mathrm{M}(D)_{x,\ y}$ ranges between 0 and 1.

Suppose that the sample size is n, and the number of grid divisions is less than $B(n)$. Then the maximal information coefficient (MIC) is defined as

$$\mathrm{MIC(D)}=\ \max\nolimits_{xy<B(n)}\left\{M(D)_{x,y}\right\}.$$

The MIC is a generalized correlation and ranges between 0 and 1. It can detect a wide range of correlations when the sample size is sufficiently large. When MIC = 0, X and Y are independent.

Maximal correlation coefficient (MCC)

MCC first performs the best conversion on two random variables X and Y, and then uses the Pearson correlation coefficient to calculate the correlation. The best transform estimation is based only on the data samples and has minimal assumptions about the data allocation and the form of the best transform. In particular, we don't need the transformation functions to come from a particular parameterized family or even monotonic. Let X, Y be random variables defined in the probability space (X, A, P) and they are randomly selected at (X, B_1) and (Y, B_2). Map $X : (X, A) \to (X, B_1)$, $Y : (Y, A) \to (Y, B_2)$ generates a subalgebra $A_1 = X^{-1}(B_1)$ and $A_2 = Y^{-1}(B_2)$ in A. P_i is a measure of P on A, $i = 1, 2$. Let function ϕ have a finite second moment $E|\phi|^2 = \int |\phi|^2 dP < \infty$, and have an inner product $(\phi_1, \phi_2) = E(\phi_1, \phi_2)$, and $L^2 = L^2(P)$ is the Hilbert space of A-measurable function ϕ; $L_i^2 = L_i^2(P)$ is the Hilbert space of A_i measurable function ϕ with finite second moment and same inner product. Define MCC between X and Y as: $\mathrm{MCC}(X_1, X_2) = \sup \{\rho(\phi_1(X_1), \phi_2(X_2))\}$, where $\rho(\cdot)$ is the Pearson correlation coefficient, and $\phi_1(X) \in L_1^2$, $\phi_2(Y) \in L_2^2$.

In principle, MCC measures the cosine of the angle between the linear subspaces of mean zero square integrable real-valued random variables. The maximal correlation is the supremum of $\rho(\phi_1(X), \phi_2(Y))$. The key is to choose the right function to get the exact value of the upper bound. Two variables are independent, $MCC(X, Y) = 0$, is equivalent to that L^{2}' s two subspaces L_i^2 $(i = 1, 2)$ are orthogonal. If the relationship between X and Y is linear, the MCC will degenerate into the Pearson correlation coefficient. At present study, we calculate the MCC by the commonly used ACE method, which can be easily obtained by R package *acepack* [45].

K shortest paths algorithm

In a directed network, K Shortest Paths Algorithm is used to find the path with the smallest weight from one starting node to the end node. Here the starting node is a genetic variant (e.g. SNP) and the end node is complex disease (e.g. AD). Given that the goal is to find the pathway with relatively large effect, the first step is to transform the between-node correlation such that we can apply the K Shortest Paths Algorithm. Let r_{ij} represent the general correlation (e.g. one of the above correlation quantities) between the two nodes M_i and M_j, and the direction is $X_i \to X_j$. Suppose there is a simple path from the SNP X to the outcome disease Y, $X \to M_1 \to M_2 \to Y$. Then, along this pathway, the effect of X on Y can be represented as $r_{X,M_1} \times r_{M_1,M_2} \times r_{M_2,Y}$. We transform r_{ij} by the following equation:

$$r_{ij}{}' = \log \frac{1}{r_{ij}},$$

where i, j represents nodes such as X, Y, and M_i $(i = 1, 2)$, then the weight along this pathway is

$$\begin{aligned} r_{X,M_1}{}' + r_{M_1,M_2}{}' + r_{M_2,Y}{}' &= \log \frac{1}{r_{X,M_1}} + \log \frac{1}{r_{M_1,M_2}} \\ &\quad + \log \frac{1}{r_{M_2,Y}} \\ &= \log \frac{1}{r_{X,M_1} \times r_{M_1,M_2} \times r_{M_2,Y}}. \end{aligned}$$

Such simple reciprocal function transforms the maximum value of the weight into its minimum, and the log transform converts the product to the summation, then the K Shortest Paths Algorithm can be easily implemented by the commonly used deviation path algorithm [34].

Definition of the pathway importance score (PIS)

It naturally defined the pathway effect to be the product of between-node connection along specific pathway, for instance, the effect of the simple pathway $X \to M_1 \to M_2 \to Y$ is defined to be $r_{X,M_1} \times r_{M_1,M_2} \times r_{M_2,Y}$. Suppose that there are totally Q pathways within one network (e.g. there are 33 pathways from APOE genotype to AD mediated by gene expression in our real data analysis), we denote q_{50} to be the median of effects of all these Q pathways, then the pathway importance score (PIS) for the i th pathway is

$$PIS_i = (the\ effect\ of\ ith\ pathway)/q_{50},$$

where $i = 1, \ldots Q$. Intuitively, for the pathway with effect greater than the median value, it should be relatively important and the PIS should be greater than one, otherwise it should be less than one. PIS provides a simple way to quantify the importance of pathway from the genetic variant to the outcome. In practice, once obtaining the rank of all pathways, one can chose those having comparable PIS with the top one for further experimental verification.

Simulation

Various simulations were conducted to assess the performance of the above six correlation measurements together with the K Shortest paths algorithm, in identifying the potential pathway from genetic variant to the disease outcome. To make the simulations more realistic, we designed the network (Fig. 3) based on the insulin signaling pathway from KEGG, we simulated that the genetic variant can affect disease mediated by gene

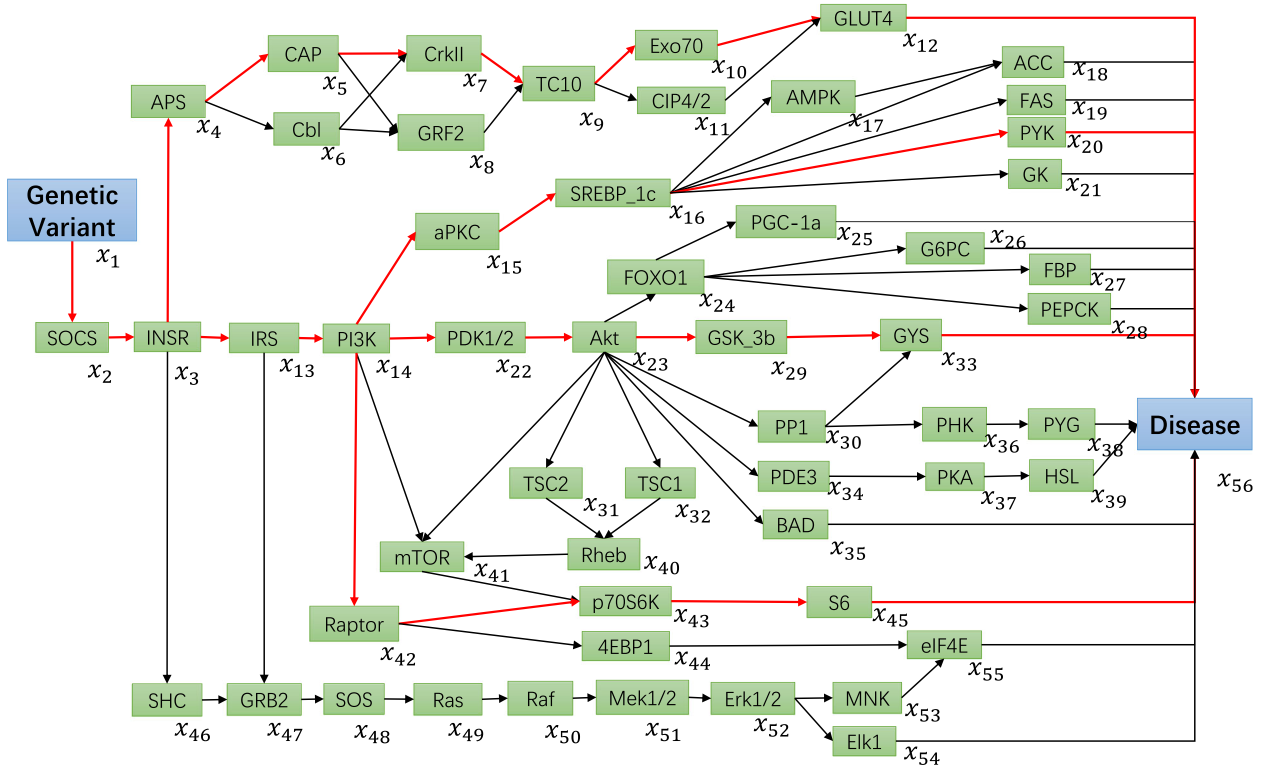

Fig. 3 The simulated network from genetic variant to disease constructed from the insulin signaling pathway from KEGG. The hypothesis is that the genetic variant can affect the disease through the gene expression on the multiple pathways. The genes and the direction are highlighted as the green box and the black arrow. The 23 edges included in the 4 effective pathways are highlighted in red

expression enriched on the insulin signaling pathway. We generated starting node X_1 from $N(0, 1)$ and then generated the other nodes following the network structure (totally 56 nodes and 82 edges). If the direction for node X_i and X_j is $X_i \rightarrow X_j$ $(i = 1, \ldots 55,\ j = 2, \ldots 56,\ i < j)$, given X_i, we generate X_j as $x_j = \mu_j + \phi(x_i) + \varepsilon_j$, where μ_j is the intercept and ε_j is the error term following $N(0, \sigma_j^2)$, the parameter μ_j and σ_j can be assigned to ensure the mean and variance of node X_j to be zero and unit. The function ϕ is pre-specified based on the designed relationship between X_i and X_j, for instance, $\phi(x_i) = \sqrt{C - x^2}$, $\phi(x_i) = x_i$, $\phi(x_i) = x_i^2$, $\phi(x_i) = \cos(x_i)$ and $\phi(x_i) = \sin(2x_i)$ for the arcuate, linear, quadratic, cosine and sine relationship, respectively. The correlation strength between these two nodes can be measured as the linear correlation coefficient between $\phi(x_i)$ and x_j, $corr(\phi(x_i), x_j)(i = 1, 2, \ldots, 56)$, which is used as the weight between these two nodes. For instance, suppose that there is the quadratic relationship between X_1 and X_2, then X_2 can be generated as $x_2 = \mu_2 + \beta_{12}x_1^2 + \varepsilon_2$, where β_{12} is the pre-specified parameter, ε_2 is the error term following $N(0, \sigma_2^2)$. We set $\sigma_2 = \sqrt{1 - 2\beta_{12}^2}$ and $\mu_2 = -\beta_{12}$. If the downstream x_j is generated from multiple upstream nodes, similar designs can be conducted. For instance, the node x_9 is linearly dependent on both x_7 and x_8, then $x_9 = \mu_j + \beta_{79}x_7 + \beta_{89}x_8 + \varepsilon_9$, and we also ensured the mean and variance of x_9 to be zero and unit respectively.

The goal is to determine the suitable correlation measure that can capture the between-node relationship in the network and can correctly pinpoint the top few pathways with large effects using the K shortest paths algorithm. Here, we assigned 4 pathways with relatively large effect, and these 4 pathways covered 23 edges (Fig. 3). Note that here we chose the correlation strength between nodes on these 4 pathways randomly from *unif*(0.75,1) and that on the other pathways randomly from *unif*(0,0.25), to make these 4 pathways have relatively larger effects than other pathways. Two scenarios are considered as follows: (1) all the between-node correlations are linear, and (2) among the 23 edges, we randomly set the proportion of the nonlinear edge (e.g. the between-node correlation is nonlinear) to be 30, 40, 50 and 60%, respectively (See Figures S16, S17, S18, S19 for details). The nonlinear relationship $(\phi(\cdot))$ includes x^2, $\cos(x), \sin(2x)$, $\sqrt{C - x^2}$. Here we used $\phi(x_i) = \sin(2x_i) + \varepsilon$ as the sine relationship, given that the sine function is close to linear in the interval $[-\pi, \pi]$. We set the proportion of nonlinear relationship to be 30% (5 edges having cosine and 2 edges having quadratic relationship), 40% (6 edges having cosine and 3 edges having quadratic relationship), 50% (8 edges having cosine and 4 edges having quadratic relationship) and 60% (8 edges having cosine and 5 edges having quadratic and 1 edge having $\sqrt{C-x^2}$ relationship). We kept the edges with the above nonlinear correlation pattern to be the same in each replicate for better comparison. For each simulation setting, we first generated the whole population with sample size 50,000, then we calculate the 82 between-node correlations (i.e. linear or nonlinear correlation between the two neighbored nodes) to derive and obtain the true order of the effects of the pathways. We randomly chose the 100, 300, 500 samples without replacement from 50,000 population, and replicated 500 simulations for each scenario.

We aimed to assess the performance of the methods of the existing six correlation metrics with K shortest paths algorithm, which can be labeled as Person-SP, Spearman-SP, DC-SP, MI-SP, MIC-SP, MCC-SP. Here we preferred two criteria to evaluate the ability of the six integration methods to pinpoint the pathways with relatively large effect. The two criteria are 1) all-right: it is the most stringent and means that the top 4 pathways can be precisely allocated with the same order as their effects from large to small; the more times to find the top 4 pathways, the better the method; and 2) range-right: it means that the top 4 pathways with some effects are ranked in top 4, while the order can be allowed to be chaotic. For instance, if the top 4 paths are sorted as $Path_1$, $Path_2$, $Path_3$ and $Path_4$ based on the pathway effect from large to small. The "all-right" criteria means that the top 4 paths must be $Path_1$, $Path_2$, $Path_3$, $Path_4$, while the "range-right" criteria means the top 4 paths just include $Path_1$, $Path_2$, $Path_3$, $Path_4$, regardless of the order. For example, it can be $Path_2$, $Path_3$, $Path_4$, $Path_1$ or any other order patterns.

Application datasets

We applied these six integration methods to identify the potential causal pathway from *APOE* genotype to AD, with the network constructed from KEGG-based Alzheimer's disease pathway (Figure S20). The Religious Orders Study and Memory and Aging Project (ROSMAP) Study is divided into two parts, ROS (The Religious Orders Study) and The Memory and Aging Project (MAP). Details about the ROSMAP can be found in previous studies [46, 47] and the website https://www.synapse.org/#!Synapse:syn3219045. In ROSMAP, Alzheimer's Disease status was determined by a computer algorithm based on cognitive test performance with a series of discrete clinical judgments made in series by a neuropsychologist and a clinician.

DNA has been used to characterize apolipoprotein E allele status (*APOE*), and more recently, it has been used to generate genome-wide genotyping data generated on a Affymetrix 6.0 platform and imputed to 2.2 million single nucleotide polymorphisms (SNPs) with HapMAP [46, 47]. The *APOE* genotype is defined as a 0–1 variable. Following previous studies [48–50], if one or both genotypes are ε4, it is assigned to be 1, otherwise it is set to be 0. The gray matter of the dorsolateral prefrontal cortex of the subject was used to extract RNA from the ROS and MAP cohorts. Agilent Bioanalyzer performs a quality assessment of samples quantified by Nanodrop. The strand-specific dUTP method [51] with poly-A se-

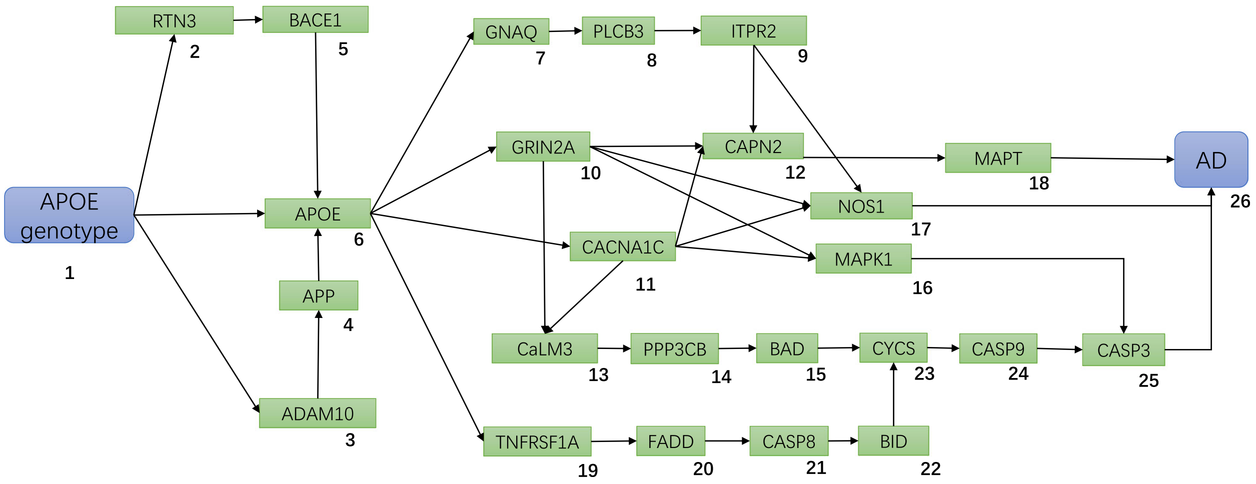

Fig. 4 The whole network from APOE genotype to AD constructed from KEGG-based Alzheimer's disease pathway. The hypothesis is that the APOE genetic variant can affect AD through the gene expression on Alzheimer's disease pathway. Multiple pathways with chain structure can be formulated with the staring node APOE genotype and the end node AD. The gene and directions are highlighted with a green frame and a black arrow

lection [52] was used in the Genomics platform of Broad Institutes for the preparation of RNA-Seq libraries. The quality of the RNA-Seq sample (Bioanalyzer RNA Integrity (RIN) score > 5) and the number threshold (5 μg) requirements were required. Sequencing was performed on the Illumina HiSeq. Before applying RSEM to estimate the expression levels of all transcripts, the non-gapped aligner Bowtie was used to compare the reads to the transcriptome reference. The result of the data RNA-Seq pipeline is the FPKM value. The quantile normalization method will first be applied to FPKM and subsequently used to eliminate potential batch effect using "*combat*" package.

The study used 364 samples (236 females and 128 males) from ROSMAP, with 163 from MAP and 201 from ROS. The age of death was between 67 and 90 years. Among these samples, 192 samples had AD and 172 were normal controls. We first performed a simple logistic regression between *APOE* genotype and AD ($\beta = 1.0595$, $p < 0.0001$). It is necessary to explore the pathway mechanism behind this association. We mapped all gene expression from ROSMAP to the KEGG Alzheimer's disease pathway to determine the candidate gene expression. The *APOE* genotype were linked to the three genes *(RTN3*, *ADAM10* and *AOPE*), which located on the left of the Alzheimer's disease pathway. Then, the other downstream genes are connected following the Alzheimer's disease pathway structure, and finally connected to AD. There is a total of 24 genes and 26 nodes in the network. The starting node of the whole network is the *APOE* genotype and the terminating node is AD (Fig. 4).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12863-020-00899-3.

Additional file 1: Table S1. The rank of the total 33 pathways for 6 integration methods. **Figure S1.** The proportion correctly pinpointing top 4 pathways under sample size 500 and nonlinear proportion 60%. **Figure S2.** The proportion correctly pinpointing top 4 pathways under sample size 500 and nonlinear proportion 30%. **Figure S3.** The proportion correctly pinpointing top 4 pathways under sample size 500 and nonlinear pattern being $\varphi(x_i) = \sin(2x_i) + \varepsilon$. **Figure S4.** The proportion correctly pinpointing top 4 pathways under sample size 500 and nonlinear pattern being exponential and reciprocal. **Figure S5. T**he proportion correctly pinpointing top 4 pathways under sample size 500 and nonlinear proportion 40% (including NPR). **Figure S6.** The proportion correctly pinpointing top 4 pathways under sample size 300 and nonlinear proportion 30%. **Figure S7.** The proportion correctly pinpointing top 4 pathways under sample size 300 and nonlinear proportion 40%. **Figure S8.** The proportion correctly pinpointing top 4 pathways under sample size 300 and nonlinear proportion 50%. **Figure S9.** The proportion correctly pinpointing top 4 pathways under sample size 300 and nonlinear proportion 60%. **Figure S10.** The proportion correctly pinpointing top 4 pathways under sample size 300 and nonlinear pattern being $\varphi(x_i) = \sin(2x_i) + \varepsilon$. **Figure S11.** The proportion correctly pinpointing top 4 pathways under sample size 100 and nonlinear proportion 30%. **Figure S12.** The proportion correctly pinpointing top 4 pathways under sample size 100 and nonlinear proportion 40%. **Figure S13.** The proportion correctly pinpointing top 4 pathways under sample size 100 and nonlinear proportion 50%. **Figure S14.** The proportion correctly pinpointing top 4 pathways under sample size 100 and nonlinear proportion 60%. **Figure S15.** The proportion correctly pinpointing top 4 pathways under sample size 100 and nonlinear pattern being $(x_i) = \sin(2x_i) + \varepsilon$. **Figure S16.** The network when there are 30% nonlinear between-node connection in the 4 effective pathways. **Figure S17.** The network when there are 40% nonlinear between-node connection in the 4 effective pathways. **Figure S18.** The network when there are 50% nonlinear between-node connection in the 4 effective pathways. **Figure S19.** The network when there are 60% nonlinear between-node connection in the 4 effective pathways. **Figure S20.** The Alzheimer's disease pathway downloaded from KEGG.

Abbreviations

GWAS: Genome-wide association studies; SNP: Single nucleotide polymorphism; MCC: The maximal correlation coefficient; PIS: The pathway importance score; AD: Alzheimer's disease; eQTLs: Expression quantitative trait loci; TWAS: Transcriptome-wide association studies; CaMMEL: The causal multivariate mediation within extended linkage disequilibrium; ROSMAP: The Religious Orders Study and the Memory and Aging Project.; MI: Mutual information; MIC: Maximal Information Coefficient; CNS: The central nervous system; LDLR: Low-density lipoprotein receptor

Acknowledgements

The results published here are based on data obtained from the AMP-AD Knowledge Portal (doi:10.7303/syn2580853). We are grateful to the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago for providing their data.

Authors' contributions

ZY conceived the study. YZ, JJ, WL contributed to data analysis. ML, LL, FX and XL contributed to the data interpretation. ZY, YZ, and JJ wrote the manuscript with help from HZ and XZ. All authors have read and approved the manuscript.

Funding

This work has been supported by grants from National Natural Science Foundation of China (81673272, 81872712, 81573259, 81803336), the Natural Science Foundation of Shandong Province (ZR2019ZD02, ZR2018BH033), and the Young Scholars Program of Shandong University (2016WLJH23). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets analyzed for this study can be found in the ROSMAP (https://www.synapse.org/#!Synapse:syn3219045). The code of MCC-SP is freely available at GitHub (https://github.com/zhuyuchen95/ADnet).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, Shandong, China. [2]Department of Data Science, School of Statistics, Shandong University of Finance and Economics, Jinan 250014, China. [3]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA. [4]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA.

Received: 22 April 2020 Accepted: 13 August 2020
Published online: 26 August 2020

References

1. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101(1):5–22.
2. Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45(12):1452–8.
3. Ramanan VK, Risacher SL, Nho K, et al. APOE and BCHE as modulators of cerebral amyloid deposition: a florbetapir PET genome-wide association study. Mol Psychiatry. 2014;19(3):351–7.
4. Jun G, Vardarajan BN, Buros J, et al. Comprehensive search for Alzheimer disease susceptibility loci in the APOE region. Arch Neurol. 2012;69(10):1270–9.
5. Grupe A, Abraham R, Li Y, et al. Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Hum Mol Genet. 2007;16(8):865–73.
6. Gottschalk WK, Mihovilovic M, Roses AD, Chiba-Falek O. The role of upregulated APOE in Alzheimer's disease etiology. J Alzheimers Dis Parkinsonism. 2016;6(1):209. https://doi.org/10.4172/2161-0460.1000209.
7. Fernandez CG, Hamby ME, McReynolds ML, Ray WJ. The role of APOE4 in disrupting the homeostatic gunctions of astrocytes and microglia in aging and Alzheimer's disease. Front Aging Neurosci. 2019;11:14. https://doi.org/10.3389/fnagi.2019.00014.
8. Xu Q, Bernardo A, Walker D, Kanegawa T, Mahley RW, Huang Y. Profile and regulation of Apolipoprotein E (ApoE) expression in the CNS in mice with targeting of green fluorescent protein gene to the ApoE locus. J Neurosci. 2006;26(19):4985–94.
9. Parcon PA, Balasubramaniam M, Ayyadevara S, et al. Apolipoprotein E4 inhibits autophagy gene products through direct, specific binding to CLEAR motifs. Alzheimers Dement J Alzheimers Assoc. 2018;14(2):230–42.
10. Lambert J-C, Berr C, Pasquier F, et al. Pronounced impact of Th1/E47Cs mutation compared with −491 at mutation on neural APOE gene expression and risk of developing Alzheimer's disease. Hum Mol Genet. 1998;7(9):1511–156.
11. Raber J, Huang Y, Ashford JW. ApoE genotype accounts for the vast majority of AD risk and AD pathology. Neurobiol Aging. 2004;25(5):641–50.
12. Namba Y, Tomonaga M, Kawasaki H, Otomo E, Ikeda K. Apolipoprotein E immunoreactivity in cerebral amyloid deposits and neurofibrillary tangles in Alzheimer's disease and kuru plaque amyloid in Creutzfeldt-Jakob disease. Brain Res. 1991;541(1):163–6.
13. Strittmatter WJ, Saunders AM, Schmechel D, et al. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proc Natl Acad Sci. 1993;90(5):1977–81.
14. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190–5.
15. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47(11):1228–35.
16. Wu Y, Zheng Z, Visscher PM, Yang J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. Genome Biol. 2017;18(1):86.
17. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet. 2009;10(3):184–94.
18. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6(4):e1000888.
19. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015;16(4):197–212.
20. Huang Y-T, Vanderweele TJ, Lin X. Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. Ann Appl Stat. 2014;8(1):352–76.
21. Huang Y-T, Liang L, Moffatt MF, Cookson WOCM, Lin X. iGWAS: integrative genome-wide association studies of genetic and genomic data for disease susceptibility using mediation analysis. Genet Epidemiol. 2015;39(5):347–56.
22. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091–8.
23. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48(3):245–52.
24. Yuan Z, Zhu H, Zeng P, et al. Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. Nat Commun. 2020;11(1):3861.
25. Park Y, Sarkar AK, He L, Davila-Velderrain J, Jager PLD, Kellis M. A Bayesian approach to mediation analysis predicts 206 causal target genes in Alzheimer's disease. bioRxiv. Published online December 1, 2017:219428.
26. Wei Z, Li H. Nonparametric pathway-based regression models for analysis of genomic data. Biostat Oxf Engl. 2007;8(2):265–84.
27. Yao DW, O'Connor LJ, Price AL, Gusev A. Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat Genet. 2020;52(6):626–33.
28. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet. 2007;81(6):1278–83.
29. Bedelbaeva K, Snyder A, Gourevitch D, et al. Lack of p21 expression links cell cycle control and appendage regeneration in mice. Proc Natl Acad Sci U S A. 2010;107(13):5845–50.
30. Yuan Z, Ji J, Zhang T, et al. A novel chi-square statistic for detecting group differences between pathways in systems epidemiology. Stat Med. 2016; 35(29):5512–24.
31. Ji J, Yuan Z, Zhang X, et al. Detection for pathway effect contributing to disease in systems epidemiology with a case-control design. BMJ Open. 2015;5(1):e006721.
32. Hoffman W, Pavley R. A method for the solution of the $N$th best path problem. J Assoc Comput Mach. 1959;6:506–14.
33. Hershberger J, Maxel M, Suri S. Finding the k shortest simple paths: a new algorithm and its implementation. Acm Trans Algorithms. 2007;3(4):45.
34. Yen JY. Finding the K shortest Loopless paths in a network. Manag Sci. 1971;17(11):712–6.
35. Liu Z, Lin X. Multiple phenotype association tests using summary statistics in genome-wide association studies. Biometrics. 2018;74(1):165–75.
36. Ray D, Boehnke M. Methods for meta-analysis of multiple traits using GWAS summary statistics. Genet Epidemiol. 2018;42(2):134–45.
37. Gozdz A, Habas A, Jaworski J, et al. Role of N-methyl-d-aspartate receptors in the Neuroprotective activation of extracellular signal-regulated kinase 1/2 by Cisplatin. J Biol Chem. 2003;278(44):43663–71.
38. Yong S-M, Lim M-L, Low C-M, Wong B-S. Reduced neuronal signaling in the ageing apolipoprotein-E4 targeted replacement female mice. Sci Rep. 2014;4:6580.
39. Hoe H-S, Harris DC, Rebeck GW. Multiple pathways of apolipoprotein E signaling in primary neurons. J Neurochem. 2005;93(1):145–55.
40. Wu H-Y, Yuen EY, Lu Y-F, et al. Regulation of N-methyl-D-aspartate receptors by Calpain in cortical neurons. J Biol Chem. 2005;280(22):21588–93.
41. Ferreira A. Calpain dysregulation in Alzheimer's disease. ISRN Biochem. 2012; 2012:728571.
42. Garg S, Timm T, Mandelkow E-M, Mandelkow E, Wang Y. Cleavage of tau by calpain in Alzheimer's disease: the quest for the toxic 17 kD fragment. Neurobiol Aging. 2011;32(1):1–14.
43. Maher A, El-Sayed NS-E, Breitinger H-G, Gad MZ. Overexpression of NMDA R2B in an inflammatory model of Alzheimer's disease: modulation by NOS inhibitors. Brain Res Bull. 2014;109:109–16.
44. Reshef DN, Reshef YA, Finucane HK, et al. Detecting novel associations in large data sets. Science. 2011;334(6062):1518–24.
45. Breiman L, Friedman JH. Estimating optimal transformations for multiple regression and correlation. J Am Stat Assoc. 1985;80(391):580–98.
46. Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS. Overview and findings from the religious orders study. Curr Alzheimer Res. 2012;9(6):628–45.
47. Bennett DA, Schneider JA, Buchman AS, Barnes LL, Boyle PA, Wilson RS. Overview and findings from the rush memory and aging project. Curr Alzheimer Res. 2012;9(6):646–63.
48. Ossenkoppele R, van der Flier WM, Zwan MD, et al. Differential effect of APOE genotype on amyloid load and glucose metabolism in AD dementia. Neurology. 2013;80(4):359–65.
49. Gomez-Isla T, West HL, Rebeck GW, et al. Clinical and pathological correlates of apolipoprotein E ε4 in Alzheimer's disease. Ann Neurol. 1996;39(1):62–70.
50. Saunders AM, Strittmatter WJ, Schmechel D, et al. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology. 1993;43(8):1467–72.
51. Levin JZ, Yassour M, Adiconis X, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. 2010;7(9):709–15.
52. Adiconis X, Borges-Rivera D, Satija R, et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods. 2013;10(7):623–9.

Publisher's Note