

基于主成分分析、去稀疏化 LASSO、柯西 P -value 整合方法研究基因甲基化在儿童经历虐待与成年罹患抑郁障碍间的中介效应

黄煜¹ 赵良渊² 高宇钊³ 郭鹏巍⁴ 张锡敏³ 赵晋芳^{5△}

【摘要】目的 识别介导儿童期经历中等至极端虐待与抑郁障碍相关的甲基化基因,有助于在基因水平上阐明抑郁障碍的发病机制,同时也可作为抑郁障碍的临床治疗提供靶点。**方法** 本研究将经典主成分分析、去稀疏化 LASSO、柯西 P -value 整合有机地结合起来识别具有中介效应的甲基化基因。数据来源于 Gene Expression Omnibus(GEO)数据库,分析采用基于 R 4.2.2 的高维推断(hdi)包进行分析。**结果** 采用主成分分析提取基因区域信息,可获得 18203 个基因。通过构建的高维中介效应模型对该基因数据进行分析,识别出具有潜在中介效应的 1 个甲基化基因,即基因 *ELAC1*。**结论** 采用整合的中介效应模型识别出甲基化基因 *ELAC1*,该基因在人脑中有较高的表达,可能与人类精神疾病相关,故该甲基化基因很可能介导了儿童期经历中等至极端虐待后导致成年抑郁障碍的发生。

【关键词】 主成分 中介分析 P -value 整合 去稀疏化 LASSO 甲基化

【中图分类号】 R195.1 **【文献标识码】** A **DOI** 10.11783/j.issn.1002-3674.2024.06.003

Based on Principal Component Analysis, De-sparsified LASSO, Cauchy P -value Combination Method to Study the Mediation Effects of Methylated Genes between Childhood Maltreatment and Adult Depression

Huang Yu, Zhao Liangyuan, Gao Yuzhao, et al (Department of Computer Teaching Center, Shanxi Medical University, Taiyuan 030001)

【Abstract】Objective The study is aimed at identifying the methylation genes that can mediate from childhood maltreatment to depression and can also provide a target for clinical treatment of depression. **Methods** The classical principal component analysis, de-sparsified LASSO, and Cauchy P -value combination were integrated to identify methylated genes with mediating effects. The data were downloaded from Gene Expression Omnibus(GEO) database, and the analysis was carried out with high-dimensional inference(hdi) package based on R4.2.2. **Results** The principal component analysis was used to extract gene information, and 18203 genes could be obtained from sites data. The integrated model was used to do the mediation analysis, and a methylated gene with a potential mediation effect was identified, namely the gene *ELAC1*. **Conclusion** The integrated mediation effect model identified the methylated gene *ELAC1*, which is highly expressed in the human brain and may be associated with human psychiatric disorders. Therefore, the methylated gene is likely to mediate the occurrence of adult depression after childhood maltreatment.

【Key words】 Principal components; Mediation analysis; P -value combination; De-sparsified LASSO; Methylation

抑郁障碍是最常见的精神疾病,是一类由各种原因引起的显著而持久的心境低落为主要临床特征的心境障碍。据世界卫生组织报告的数据显示,全球有超过 3.5 亿人受到抑郁症的困扰,且近十年来罹患该疾病的人数增长了约 18%,故对抑郁障碍的发病机制进行研究具有重要的意义^[1]。目前普遍认为抑郁障碍发病与遗传、心理、社会环境等多种因素有关。近年来,越来越多的研究者在表观遗传学领域对该疾病的发病机制进行了研究,且有研究结果表明脱氧核糖核酸甲基化(DNA methylation, DNAm)与抑郁障碍的发病存在一定的关联^[2]。

DNAm 作为一种机体不可或缺的基因表观遗传

修饰机制,它虽不改变 DNA 序列,但却能调节 DNA 序列的生物学活性,从而导致机体产生一些未知的生物学效应。有研究表明,儿童期经历过虐待可导致 DNAm^[3],此外,也有研究表明 DNAm 与成年罹患抑郁障碍有关^[2,4],识别 DNAm 基因对于在基因层面阐明抑郁障碍具有重要的意义。

本课题组前期研究基于位点水平,识别了介导儿童期遭受虐待与成年后罹患抑郁障碍的 DNAm 位点^[5]。但是从生物学角度考量,基因是遗传的基本功能单位;其次,基因相比于位点稳定性更好,不同的地域或人种间位点可能存在变异,但基因却一致,识别得到的结果更具有一般性;最后,从位点水平上升至基因水平,也能在一定程度上进行降维^[6]。

目前有研究表明,儿童期遭受虐待会增加成年后罹患抑郁障碍的风险^[3-4]。即在宏观层面上可发现,儿童期遭受虐待与成年后罹患抑郁障碍存在一定的关联性,且基因甲基化参与了该过程,即基因甲基化可视

1.山西医科大学计算机教学部(030001)

2.山西医科大学党委办公室、校长办公室

3.山西财经大学统计学院

4.山西财经大学财政与公共经济学院

5.山西医科大学公共卫生学院

△通信作者:赵晋芳,E-mail:zhaojinfangcn@163.com

为中介因素,介导了儿童期经历过虐待后在成年罹患抑郁障碍的发病过程。本研究基于 Beron 与 Kenny 在 1986 年提出的中介分析框架,识别具有潜在中介效应的甲基化基因,为该疾病机制的理解提供新视角^[7]。

最基础的中介效应模型为单中介模型,之后又扩展为多中介模型以及目前尚处在摸索阶段的高维中介模型^[7-8]。同时,也有研究考量了暴露与中介变量间的相互作用的研究^[9]。在我国医学研究领域,多为中介模型的应用,比如杨春艳等探讨了医学研究领域的中介分析方法的运用^[10],饶悦等采用中介分析的视角探讨了高血压患者心理灵活性和自我管理在人格特征与睡眠质量中发挥的作用^[11],孟迪等分析了疲乏在乳腺癌术后化疗患者日常生活记忆与睡眠质量间的中介作用^[12]。较少对中介模型进行方法学应用的考量。但是,这些研究选择的中介变量间多为单个变量,本研究考虑从基因水平入手,中介变量则是一个基因区域,如果想引入该基因区域进入模型,则首先需要聚集基因区域内的位点信息。

基于以上的考虑,本研究将经典主成分分析(principal components analysis, PCA)、Wilk's Lambda 检验、去稀疏化最小绝对值压缩选择算子(LASSO)以及柯西 P 值整合(P-value)有机地结合起来进行中介分析。该模型虽整合了多种方法,但仍然处于基本的中介分析框架之中,将基因作为中介变量引入模型后,只能识别出具有潜在中介效应的基因,无法估计其中介效应的大小。

资料与方法

1. 资料来源

研究数据来自于美国格雷迪创伤项目研究,该研究在获得埃默里大学医学院和格雷迪纪念医院机构审查委员会批准后,将亚特兰大市中心居民中招募的非裔美国人作为研究对象。该数据集包含基线信息、细胞成分资料和 DNAm 数据。研究采用改良的贝克抑郁量表(Beck depression inventory, BDI)对研究对象进行测量^[13],且筛选出 BDI 评分 ≥ 14 的观察单位。在观察结果中,排除了接受过抑郁障碍治疗的研究对象,因为接受的治疗可能会影响基因甲基化的改变,从而使中介效应分析的结果变得更加难以解释。同时,研究也剔除了儿童期经历中度至极端虐待缺失的观察单位,最后纳入中介效应分析的研究对象有 138 例。关于该研究项目的详细信息,读者可以在网站 <http://gradytraumaproject.com/> 进行参阅。研究采用的 DNAm 数据、临床数据和细胞数据可以在以下网址进行下载获取 <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72680>。

2. 方法

(1) 信息提取

本研究使用经典 PCA 从每个基因区域中提取主要信息。首先,根据基因包含的位点数对基因进行筛选,本研究对基因位点数 ≥ 4 的基因提取主成分,保留基因位点数 ≤ 3 的基因。最后,获得了一个能反映基因整合信息的数据,该数据中部分是主成分(PC_s),部分是原始数据。

(2) 变量筛选

该步采用 Wilk's lambda 检验筛选出一定数量的具有最大效应的基因^[14]。该方法采用的策略与 Fan 等在 2008 年提出的确定性独立筛选(sure independence screening, SIS)方法类似^[15],主要区别在于本研究筛选的是一定数量具有最大效应的基因。本研究保留的基因数量取决于因变量 Y 的类型。当 Y 是连续

变量时,选择 $d = \lceil \frac{2n}{\log(n)} \rceil$ 个在模型中效应最大基因。

对于 Y 是二分类变量时,保留最大效应的基因个数减小至 Y 为连续变量时的 1/4,即 $d = \lceil n / (2 \log(n)) \rceil$ 。该检验的主要目的是获得 P-values, P-values 的大小反映了从 X 到 PC_i, $i = 1, \dots, k$ 效应的强弱。P-values

可以用 $P_{i,a} = \left\{ 1 - F_{K_i, n-K_i-1} \left(\frac{(n-K_i-1)(1-\Lambda_i)}{K_i \Lambda_i} \right) \right\}$ 来计算,其中 $\Lambda_i = \frac{|W_i|}{|B_i + W_i|}$ 是 Wilk's Lambda 统计量, $W_i =$

$\sum_{h=1}^2 n_h (\hat{a}_{ih} - \hat{\theta}_{2i}) (\hat{a}_{ih} - \hat{\theta}_{2i})^T$, $B_i = \sum_{h=1}^2 \sum_{j=1}^{n_h} (PC_{ij} - \hat{a}_{ih}) (PC_{ij} - \hat{a}_{ih})^T$ 。筛选后,基因的数量从 k 下降到 d,研究只纳入这 d 个基因,并检验它们的中介作用。将筛选出的基因表示为 $[PC]_i, i = 1, \dots, d$ 。

(3) 利用去稀疏化 LASSO 来得到 P-values

本研究先拟合下面的模型(1),

本研究先拟合下面的模型(1),

$$E(Y) = \theta_3 + c_3 X + \sum_{i=1}^d b_i PC_{iK_i} \quad (1)$$

经过筛选后,维数 $D = \sum_{i=1}^d K_i$ 仍可能很大,最小二乘估计可能仍无法进行估计。我们同样使用 \hat{b}_{Lasso} 表示 LASSO 参数估计值。对于连续型因变量 Y,我们得到一个偏差校正估计量,可以表示如下:

$$\hat{b}_j = \frac{Z_j^T Y}{Z_j^T PC_j} - \sum_{l \neq j} \frac{Z_j^T PC_l}{Z_j^T PC_j} \hat{b}_{Lasso,l} \quad (2)$$

式中, \hat{b}_j 是第 j 个 PC_j 的偏差校正系数; $\hat{b}_{Lasso,l}$ 是通过拟合 LASSO 回归估计出的第 l 个 PC_l 的系数; Z_j 是由 $Z_j = PC_j - PC_{-j} \hat{\gamma}_{Lasso}$ 获得的正则化残差,其中 $\hat{\gamma}_{Lasso}$ 是基于 LASSO 回归得到的回归系数,若用 PC_{-j} 表示除 PC_j 之外的所有变量,那么通过 PC_j 与 PC_{-j} 进行 LASSO 回归,即可获得该回归系数。Van de Geer 等在 2014 年证明了去稀疏化估计的渐近正态性^[16],

$$u_i = \frac{\sqrt{n}(\hat{b}_i - \gamma_i^0)}{\sigma_\epsilon \sqrt{\Omega_{ii}}} \xrightarrow{d} N(0, 1) \text{ as } p \geq n \rightarrow \infty$$

其中 γ_i^0 表示真实的回归系数, σ_ϵ 可通过标准化 LASSO 回归计算得到, Ω_{ii} 可通过以下公式(3)计算得到,

$$\Omega_{ii} = \frac{nZ_i^T Z_i}{[Z_i^T Z_i][Z_i^T Z_i]} \quad (3)$$

在零假设成立的条件下, $H_0: \gamma_i^0 = 0$, 我们利用渐近正态性原理得到关于 D 个 PC_s 的 P -values。

当研究的结局变量为二分类变量时, Van de Geer 同样证明了去稀疏化估计的渐近正态性^[16]。设定 $W = (X, M)^T, \beta = (c', b)^T$ 以及损失函数 $L_\beta(y, W) = L(y, W\beta)$; 同时再定义 $\dot{L}_\beta = \frac{\partial}{\partial \beta} L_\beta, \ddot{L}_\beta = \frac{\partial^2}{\partial \beta \partial \beta^T} L_\beta$ 以及 $\varphi_L = \sum_{i=1}^n L(y_i, w_i^T \beta) / n$ 。利用 LASSO 回归对中介效应进行估计, 可得 $\hat{\beta} = \underset{\beta}{\operatorname{argmin}} (\varphi_L + \lambda \|\beta\|_1)$, 其中 λ 为调和参数。

我们定义 $\hat{\Sigma} = \varphi_{\ddot{L}_\beta}$ 以及通过逐节点 LASSO 构建 $\hat{\Theta} = \hat{\Theta}_{Lasso}$ 。此时, 可获得去稀疏化 LASSO 估计 $\tilde{\beta} = \hat{\beta} - \hat{\Theta} \varphi_{\dot{L}_\beta}$ 。Van de Geer 给出了在广义线性模型框架下去稀疏化 LASSO 参数估计的算法; 同时也证明了去稀疏化估计的渐近正态性, 即

$$u_i = \frac{\sqrt{n}(\tilde{\beta}_i - \gamma_i^0)}{\hat{\sigma}_i} \xrightarrow{d} N(0, 1) \text{ as } p \geq n \rightarrow \infty$$

其中, $\hat{\sigma}_i^2 = (\hat{\Theta} P_{\dot{L}_\beta \dot{L}_\beta} \hat{\Theta}^T)_{i,i}$ 。此时, 我们基于渐近正态性原理可得到每个 PC 的 P -value。

我们采用 $P_s = (P_{1,s}, P_{2,s}, \dots, P_{D,s})$ 表示 D 个 PC 的 P -values; 当 Y 为连续型变量, 可以利用 $P_{j,s} = 2$

$\left\{ 1 - \Phi\left(\frac{\sqrt{n}\hat{b}_j}{\sigma_\epsilon \sqrt{\Omega_{jj}}}\right) \right\}$ 来获得 P -value; 当 Y 为离散型变量, 可通过 $P_{j,s} = 2 \left\{ 1 - \Phi\left(\frac{\sqrt{n}\hat{b}_j}{\hat{\sigma}_j}\right) \right\}$ 来计算 P -value, 其中

量, 可通过 $P_{j,s} = 2 \left\{ 1 - \Phi\left(\frac{\sqrt{n}\hat{b}_j}{\hat{\sigma}_j}\right) \right\}$ 来计算 P -value, 其中

$$\Omega_{jj} = \frac{nZ_j^T Z_j}{[Z_j^T Z_j][Z_j^T Z_j]}, \hat{\sigma}_j^2 = (\hat{\Theta} P_{\dot{L}_\beta \dot{L}_\beta} \hat{\Theta}^T)_{j,j}$$

(4) 采用柯西 P -value 整合方法获得各基因的 P -values

本研究利用 $P_g = (P_{1,g}, P_{2,g}, \dots, P_{d,g})$ 表示 d 个基因的 P -values, 利用 $\{P\}_{K_j}$ 表示第 j 个基因中的 PCs 对应的 P -value。 P_g 中每个元素都是通过整合对应的 $\{P\}_{K_j}$ 得到 P -value。

$$\{P\}_{K_j} = \begin{cases} (P_{1,s}, P_{2,s}, \dots, P_{K_j,s}), j=1 \\ (P_{\sum_{t=1}^{K_j} K_j+1,s}, P_{\sum_{t=1}^{K_j} K_j+2,s}, \dots, P_{\sum_{t=1}^{K_j} K_j,s}), j=2, 3, \dots, d \end{cases} \quad (4)$$

本研究中采用基于柯西分布的 P -value 整合方法, Liu 等于 2018 提出该 P -value 整合方法, 且该方法可对具有任意结构的 P -value 进行整合^[17]。在本研究中, 柯西 P -value 整合检验统计量及其对应的 P -value 分别表示为 $T_{j,g}$ 和 $P_{j,g,cauchy}$ 。

$$T_{j,g} = \begin{cases} \sum_{t=1}^{K_j} \frac{1}{K_j} \tan\{(0.5 - \{P\}_{K_j,t})\pi\}, j=1 \\ \sum_{t=\sum_{j=1}^{j-1} K_j+1}^{K_j} \frac{1}{K_j} \tan\{(0.5 - \{P\}_{K_j,t})\pi\}, j=2, 3, \dots, d \end{cases} \quad (5)$$

$$P_{j,g,cauchy} = 0.5 - \frac{\tan^{-1}(T_{j,g})}{\pi}, j=1, 2, \dots, d \quad (6)$$

(5) 联合显著性检验

采用联合显著性检验来判断检验的基因是否具有潜在的中介效应, 本研究中只有当 $P_{i,a}$ 和 $P_{i,g}$ 都显著时, 才拒绝不存在中介效应的零假设, 此时才认为该基因具有中介效应。联合显著性检验的 P -value 定义为 $P_{i,M} = \operatorname{Max}(P_{i,a}, P_{i,g})$, 检验水准 $\alpha = 0.05$ 。只有当 $P_{i,M} < \alpha$ 时, 我们才认为检验的基因具有中介效应。模型构建的步骤如图 1 所示。

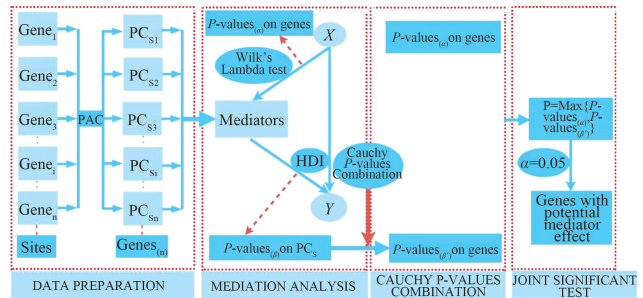


图 1 高维中介效应模型构建步骤

结果

本研究采用整合的中介效应模型来探究儿童期经历中等至极端虐待与抑郁障碍之间的关系, 并进一步评估基因甲基化在这个过程中发挥的中介效应作用。在利用 DNAm 数据进行分析之前, 首先要将 DNAm 数据中表示甲基化程度的 β 值转化为 M 值, 即 $M = \log\left(\frac{\beta}{1-\beta}\right)$ ^[18]。清理之后的 DNAm 数据包含了 335669 个位点, 本研究基于基因水平, 需要进行位点与基因之间的映射, 最后获得 18203 个基因。

(1) 信息提取

根据基因区域内包含的位点数进行筛选, 基因位点数 ≥ 4 的基因为 15029 个, 基因位点数 ≤ 3 的基因为 3174 个。对于基因区域内位点数 ≥ 4 的 15029 个基因提取主成分, 且根据累计变异解释度要达到 80% 确定每个基因区域内的主成分个数。

(2) 基因筛选

由于 DNAm 数据的维数过高, 在采用模型分析之

前,采用与 Fan 等提出的策略类似的 Wilk's lambda 检验进行筛选。筛选保留的甲基化基因个数,不仅与样本含量有关,同时也与结局变量的类型有关,设定样本含量为 n ,当结局变量为分类变量时,筛选保留的 DNAm 位点个数为 $\lceil \frac{n}{2\log(n)} \rceil$,当结局变量为连续型变量时,筛选保留的位点个数为分类变量的 4 倍,即 $\lceil \frac{2n}{\log(n)} \rceil$ 。由于研究采用的样本含量 $n = 138$,且结局变量为 BDI 评分,为连续型变量,根据该方法筛选出保留的甲基化基因个数为 $\lceil \frac{2n}{\log(n)} \rceil = 57$ 。中介因素为

$$\begin{cases} BDI = \theta_1 + \beta_{a1_maltreatment} Maltreatment + \varepsilon_1 \\ GENE_m_j = \theta_{2j} + \alpha_{a2j_maltreatment} Maltreatment + \beta_{c2j}^T CellType + \varepsilon_{2j}, j = 1, 2, \dots, K \\ BDI = \theta_3 + \beta_{a3_maltreatment} Maltreatment + \sum_{j=1}^K \beta_{mj} GENE_m_j + \beta_{c3}^T CellType + \varepsilon_3 \end{cases} \quad (7)$$

先采用模型(7)的第一个回归方程来确定结局变量 BDI 评分与暴露变量儿童期经历过中度至极端虐待的关系,即为 $\hat{\beta}_{a1_maltreatment} = 2.765, P = 0.0741$ 。依据检验水准 $\alpha = 0.05$,尚不能认为成年后罹患抑郁与儿童期经历过中度至极端虐待间存在统计关联,提示儿童期经历过中度至极端虐待与成年后罹患抑郁可能存在完全中介效应。本研究采用模型(7)中的第二个回归方程进行估计基因筛选,采用模型(7)最后一个回归方程进行高维推断,获得 57 个基因涵盖的 494 个 PC_s 与 18 个原始数据对应的 P -values。

(4) 采用柯西 P -value 整合方法获得 57 个基因的 P -values

采用柯西 P -value 整合方法,整合 57 个基因区域的 P -values,最终获得对应的 57 个基因的 P -values。

(5) 联合显著性检验

对同一基因在第(2)和(4)步中对应的 P -values 取最大值,并依据 $\alpha = 0.05$ 的检验水准,识别出位于 18 号染色体上具有潜在中介效应的甲基化基因 *ELAC1*。

讨 论

本研究采用整合的高维中介模型识别出了介导抑郁障碍发病的潜在中介效应基因 *ELAC1*。图 2 反映出基因 *ELAC1* 在人脑中表达较高,在人体的其他组织中也有表达。对于该基因未找到直接的研究证据支持基因 *ELAC1* 与抑郁障碍的发病有关,但是 Lee 等在苹果酸酶-2 与精神病和躁狂症的易感性的研究中提及了该基因,且与该基因连锁不平衡的任一基因内的突变仍可能是关联存在的原因,可知该基因与精神疾病有关联^[21]。该基因很可能介导了成年抑郁障碍的发生,但要获得确定性的结论,则需要进行生物学实验才能验证。

57 个甲基化基因,其中包括 494 个 PC_s, 18 个原始甲基化数据以及他们对应的 P -values。

(3) 利用去稀疏化 LASSO 来得到 P -value

研究采用 57 个甲基化基因以及 BDI 评分,构建中介模型,如图 1 所示。在构建模型时,首先,筛选与结局变量 BDI 评分相关的协变量,筛选未发现存在有统计学意义的协变量,故在构建模型时,我们只考虑暴露因素,即儿童期经历过中度至极端虐待,但考量到细胞数据对基因甲基化数据的影响,在引入基因甲基化数据时,要引入细胞成分数据来校正基因甲基化^[19-20]。最终,研究构建了中介模型(7)。

本研究采用了 Baron 与 Kenny 提出的中介模型框架^[7],并引入多种统计方法解决信息整合与高维的问题,构建了基于基因的高维中介效应模型。该方法适用于具有层级关系的高维数据,即收集的原始数据为低层级数据,但研究需要采用较高层级数据,故需要对低层级数据进行数据整合。在生物学研究领域, DNAm 数据的分析就存在这种情况,原始数据是关于位点的数据,但是考虑到位点非遗传的基本单位,且在人群中稳定性差,所以研究拓展至基于基因区域的研究,对位点信息进行有效提取即可获得基因信息。但是,利用该模型进行分析时,未涉及基因与基因间复杂的相互作用,可能是多阶交互作用,抑或是基因间存在网状结构。而该模型将所有的甲基化基因都进行独立性统一处理;该研究虽然是基于基因水平,但采用经典主成分分析提取信息,该方法默认了基因区域内位点间的关系是线性的,若基因区域内位点间存在非线性关系时,单纯采用经典主成分分析将无法有效地提取信息。同时,该模型在识别具有潜在中介效应的甲基化基因时,无法估计出甲基化基因中介效应的大小。此外,在模型拟合的过程中,尚无法控制方向不同的中介效应的相互抵消,致使可能无法识别出具有潜在中介效应但存在相互抵消的基因,识别出的甲基化基因均为具有显著性意义的甲基化基因。最后,该模型是基于全局角度构建的模型,与目前其他研究者提出的高维中介效应模型存在一定的差别。Fang 等构建的高维中介效应模型基于单个基因区域,对基因逐个进行分析研究,即将基因区域逐个纳入中介分析框架进行研究,未能考虑基因间的相互关系^[22]。

从生物学的角度考量,可知基因分为主效应基因与微效基因,对于主效应基因,可能单个位点即具有显著的生物学效应,但是对于微效基因,则单个位点的生

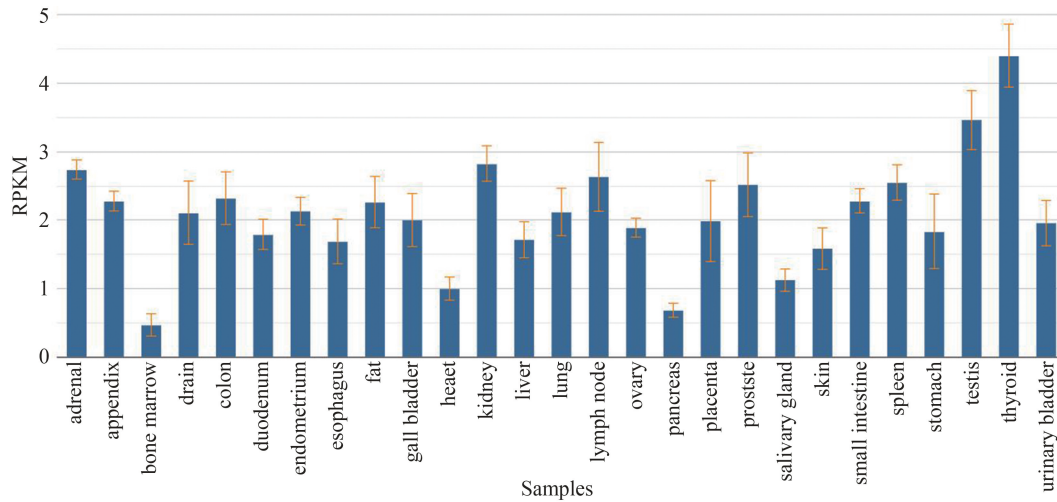


图 2 基因 *ELAC1* 在不同人体组织中的表达量

该图来自于 <https://www.ncbi.nlm.nih.gov/gene/55520>

物学效应较小,只有将这些微效应累加起来才能发挥显著的生物学效应,该整合模型整合基因区域内的信息,却只识别出一个甲基化基因,可能是基因区域内的信息提取方法存在不足,未对基因区域内位点间的非线性关系进行充分的考量,核主成分分析可以提取非线性关系,但是基于核函数的方法,提取的核主成分个数往往与样本量相等,即便是基因区域内的位点较少,提取的核主成分数也与样本数相近,从全局的角度考量,该方法不仅没有降维,反而提高了维数,存在明显的不足。如何有效地提取基因区域内位点的信息,且兼顾线性与非线性关系,将是下一阶段研究的内容。

参 考 文 献

- [1] World Health Organisation. Depression and Other Common Mental Disorders [EB/OL]. (2017) [2021-05-09]. <http://apps.who.int/iris/bitstream/10665/254610/1/WHO-MSD-MER-2017.2-eng.pdf?ua=1>.
- [2] Chen D, Meng L, Pei F, et al. A review of DNA methylation in depression [J]. *Journal of Clinical Neuroscience*, 2017, 43 : 39-46.
- [3] Suderman M, Borghol N, Pappas JJ, et al. Childhood abuse is associated with methylation of multiple loci in adult DNA [J]. *BMC Medical Genomics*, 2014, 7(1) : 13.
- [4] Styron T, Janoff-Bulman R. Childhood attachment and abuse: Long-term effects on adult attachment, depression, and conflict resolution [J]. *Child Abuse & Neglect*, 1997, 21(10) : 1015-1023.
- [5] 黄煜,孙翔飞. 基于贝叶斯方法研究 DNA 甲基化在儿童经历虐待与成年罹患抑郁障碍间的中介效应 [J]. *中国卫生统计*, 2021, 38(5) : 759-762.
- [6] Neale BM, Sham PC. The future of association studies: gene-based analysis and replication [J]. *The American Journal of Human Genetics*, 2004, 75(3) : 353-362.
- [7] Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations [J]. *Journal of Personality and Social Psychology*, 1986, 51(6) : 1173.
- [8] Zhang H, Zheng Y, Zhang Z, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies [J]. *Bioinformatics*, 2016, 32(20) : 3150-3154.
- [9] Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros [J]. *Psychological Methods*, 2013, 18(2) : 137.
- [10] 杨春艳,侯艳,李康. 中介分析方法及其在医学研究中的应用 [J]. *中国卫生统计*, 2017, 34(1) : 159-162.
- [11] 饶悦,莫祖娥,姜振英,等. 高血压患者心理灵活性,自我管理在人格特征与睡眠质量中的中介效应研究 [J]. *中国卫生统计*, 2020, 37(2) : 98-100+103.
- [12] 孟迪,崔慧霞, Fraidon Frahmand, 等. 疲乏在乳腺癌术后化疗患者日常生活记忆与睡眠质量间的中介作用分析 [J]. *中国卫生统计*, 2020, 37(4) : 84-86.
- [13] Beck AT, Ward C, Mendelson M, et al. Beck depression inventory (BDI) [J]. *Arch Gen Psychiatry*, 1961, 4(6) : 561-571.
- [14] Kanti M, John T, John B. *Multivariate Analysis* [M]. New York: Academic Press, 1979.
- [15] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space [J]. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2008, 70(5) : 849-911.
- [16] Van de Geer S, Bühlmann P, Ritov Y, et al. On asymptotically optimal confidence regions and tests for high-dimensional models [J]. *The Annals of Statistics*, 2014, 42(3) : 1166-1202.
- [17] Liu Y, Xie J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures [J]. *Journal of the American Statistical Association*, 2020, 115(529) : 393-402.
- [18] Du P, Zhang X, Huang CC, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis [J]. *BMC Bioinformatics*, 2010, 11(1) : 587.
- [19] Jones MJ, Islam SA, Edgar RD, et al. Adjusting for cell type composition in DNA methylation data using a regression-based approach [M] // *Population Epigenetics*. New York: Humana Press, 2015:99-106.
- [20] Rahmani E, Zaitlen N, Baran Y, et al. Correcting for cell-type heterogeneity in DNA methylation: a comprehensive evaluation [J]. *Nature Methods*, 2017, 14(3) : 218-219.
- [21] Lee BD, Walss-Bass C, Thompson PM, et al. Malic enzyme 2 and susceptibility to psychosis and mania [J]. *Psychiatry Research*, 2007, 150(1) : 1-11.
- [22] Fang R, Yang H, Gao Y, et al. Gene-based mediation analysis in epigenetic studies [J]. *Briefings in Bioinformatics*, 2021, 22(3) : bbaa113.

(责任编辑:张悦)