浙江理工大学学报，2020，43(5)：687-692
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673－3851(n).2020.05.014

基于GWAS的多元关联分析在精神类疾病遗传相关性分析中的应用

侯龙傲，贺平安
（浙江理工大学理学院，杭州 310018）

摘　要：基于躁郁症、重度抑郁症、精神分裂症的三种疾病的全基因组关联分析(Genome-wide association study，GWAS)数据，将基于基因的多元全基因组关联分析模型(Multivariate gene-based genome-wide association analysis，MGAS)分别应用于混合线性模型和LD评分回归模型(LD score regression model，LD Hub)提供的遗传相关性数据分析，检测得到两组与三种疾病相关的基因。结果表明：对于混合线性模型和LD Hub模型提供的遗传相关性数据，两者多元关联分析检测出的显著性基因有较强的一致性。因此，作为一种替代策略，在由原始数据计算表型相关性困难的情况下，LD Hub模型得到的遗传相关性可以应用于更多类疾病的多元关联分析。

关键词：全基因组关联分析；多元关联分析；遗传相关性；躁郁症；重度抑郁症；精神分裂症
中图分类号：O29　　**文献标志码**：A　　**文章编号**：1673-3851 (2020) 05-0687-06

Application of GWAS-based multivariate association analysis in genetic correlation analysis of psychiatric disorders

HOU Longao, *HE Pingan*
(School of Science, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Based on the Genome-wide association study (GWAS) data of the three diseases of bipolar disorder major depressive disorder, and schizophrenia, the gene-based multivariate gene-based genome-wide association analysis model (MGAS) was used to analyze the genetic correlation data provided by the mixed linear model and the LD score regression model (LD Hub), and obtain two groups of genes related to the three diseases. The results show that: for the genetic correlation data provided by the mixed linear model and LD Hub model, the significant genes detected by the multivariate association analysis have strong consistency. Thus, when it is difficult to calculate the phenotypic correlation from the original data, the genetic correlation obtained by LD Hub model as an alternative strategy can be applied for the multivariate association analysis of more types of diseases.

Key words: genome-wide association study; multivariate association analysis; genetic correlation; bipolar disorder; major depressive disorder; schizophrenia

0　引　言

目前精神类疾病的致病机制尚不清楚，因此很难确定诊断标准。Manolio等[1]的研究表明，遗传因素在所有精神疾病的主要病因中占举足轻重的地位，研究患病个体基因型与精神类疾病的表型之间的相关性，成为目前生物医学研究的重要方向。Cole等[2]的研究表明，精神类疾病存在一定的基因

收稿日期：2020－01－16　　网络出版日期：2020－04－02
基金项目：国家自然科学基金项目(61772027)
作者简介：侯龙傲(1994－　)，男，河南漯河人，硕士研究生，主要从事生物信息学方面的研究。
通信作者：贺平安，E-mail：pinganhe@zstu.edu.cn

重叠,即一个基因可能是多个精神类疾病的公共致病基因。随着精神类疾病全基因组基因型数据的快速增长,在分子水平上研究多个精神类疾病的公共致病基因成为可能,而全基因组关联分析(Genome-wide association study, GWAS)为此提供了新的思路。周家蓬等[3]的研究表明,GWAS是在人类全基因组水平上对单核苷酸多态性(Single nucleotide polymorphism, SNP)进行相关性分析,检测与表型相关的遗传因子。GWAS作为研究致病原因的主要工具,广泛应用于鉴别复杂疾病与序列变异的关联分析中。

GWAS揭示了成千上万的疾病风险位点,但仍受到两方面的限制:基因是基因组的基本功能单元;表型间往往存在相关性,如高血压、中风患者更易患阿尔兹海默症(Alzheimer disease, AD)。因此基于基因的关联分析替代位点水平关联分析、多元关联分析替代传统的单疾病分析,能更好地探究复杂疾病潜在机制。多种多元关联分析模型已应用于复杂疾病的全基因组关联分析中,如O'Brien模型[4]、多表型联合模型(Joint model of multiple phenotypes, MultiPhen)[5]、基于基因的多元全基因组关联分析模型(Multivariate gene-based genome-wide association analysis, MGAS)[6]等,这些模型均是以GWAS数据作为分析源的多元分析工具。

多元关联分析往往需要考虑表型间的遗传相关性,不同模型得到的表型间的遗传相关性不同。Lee等[7]提出的线性混合模型虽然精度较高,但由于样本、位点等数据的限制,得到疾病遗传相关性的疾病种类较少,这限制了基因水平多元关联分析的疾病选择;Bulik-Sullivan等[8]提出了跨性状LD评分回归模型(LD score regression model, LD Hub),以计算疾病间的遗传相关性,该模型仅需GWAS统计数据而且不会因为样本重叠使结果产生偏差,因此大多数精神类以及非精神类疾病的相关性结果都可以在LD Hub中查询。不同遗传相关性模型的结果应用于多元关联分析模型的可靠性是一个重要的研究课题。

MGAS模型可以对不相关的个体进行基因水平的多性状关联分析,通过对基于标准单疾病全基因组序列变异的统计关联测试软件中获得的 p 值信息排序、加权,获得基于基因的多疾病综合相关性 p 值 P_{MGAS}。MGAS模型通过模拟数据以及真实临床数据的检验,其结果具有合适的I型错误率,且由于MGAS模型不需要对表型、基因型数据排序,相较于需要排序的关联分析方法效能更高。因此,本文选择多元关联分析MGAS模型作为主要工具,分析三种精神疾病与遗传基因之间的相关性。

在躁郁症(Bipolar disorder, BIP)、重度抑郁症(Major depressive disorder, MDD)、精神分裂症(Schizophrenia, SCZ)三种疾病中,线性混合模型与LD Hub模型的遗传相关性数据[7-8]都是存在的。本文基于这三种疾病的GWAS数据,分别将线性混合模型与LD Hub模型的遗传相关性数据代入MGAS模型,计算三种疾病与对应基因的综合关联性,预测与这三种精神类疾病密切相关的公共致病基因。此外,通过比较MGAS模型在线性混合模型与LD Hub模型的遗传相关性上的计算结果,分析这两类遗传相关性数据对MGAS模型的影响。如果结果是一致的,则当线性混合模型中没有疾病遗传相关性数据时,可以使用LD Hub模型的数据计算疾病与基因的综合关联性。

1 数据与分析方法

1.1 数据下载

本文以躁郁症、重度抑郁症、精神分裂症这三种常见的被认为相关性较强的精神类疾病作为分析对象。为了保证与位点间连锁不平衡(Linkage disequilibrium, LD)的一致性,这三种疾病的样本数据、祖先LD信息数据均来自欧洲样本。这三种疾病的GWAS荟萃分析结果可从精神病基因组学联盟(https://www.med.unc.edu/pgc/)下载,精神分裂症GWAS数据的3个位点相关信息样例见表1。

表1 精神分裂症的GWAS数据

SNP	CHR	BP	A_1	A_2	OR	SE	PVAL	INFO	NGT
rs17693963	6	27818144	A	C	1.2530	0.0352	1.56×10^{-10}	1.0055	4
rs7746199	6	27261324	T	C	0.8414	0.0275	3.49×10^{-10}	1.0020	14
rs13194781	6	27815639	A	G	1.2554	0.0367	5.45×10^{-10}	1.0054	4

注:SNP表示位点名;CHR表示染色体;BP表示位点位置;A_1表示参考等位基因;A_2表示影响等位基因;OR表示回归优势比;PVAL表示位点 p 值;INFO表示方差比;NGT表示有关该位点的研究数;SE表示比值比标准差。

本文将分析重点放在最小等位基因频率不小于 0.01 且估算质量分数大于 0.6 的常染色体位点上。本文使用的位点水平 p 值能减少系统偏差并尽可能降低假阳性结果的可能性。躁郁症数据包含了 2427220 个常染色体位点，样本包括 7481 名躁郁症受试者和 9250 名对照组样本；重度抑郁症数据包含了 9533408 个常染色体位点，样本包括 16823 名重度抑郁症受试者和 25632 名对照组样本；精神分裂症数据包含 1252901 个常染色体位点，样本包括 34241 名精神分裂症受试者和 45604 名对照组样本。

1.2　数据预处理

考虑到 Ward 等[9] 关于非编码变体在复杂性状和疾病中的潜在重要作用的研究，以及 Gamazon 等[10] 对非编码变体进行基于基因的研究，测试中的输入文件也包括了基因中非编码位点。汇总数据的修剪处理后，这三种疾病包含了 742962 个公共常染色位点（共有 742962 个位点与这三种疾病都有 p 值数据，输入数据的前 3 行见表 2）。

表 2　位点与三种疾病 p 值的输入数据（前 3 行）

SNP	CHR	BP	BIP	MDD	SCZ
rs17693963	6	27818144	1.86×10^{-3}	1.24×10^{-10}	1.56×10^{-10}
rs7746199	6	27261324	1.61×10^{-3}	2.24×10^{-10}	3.49×10^{-10}
rs13194781	6	27815639	3.92×10^{-3}	1.19×10^{-7}	5.45×10^{-10}

注：SNP 表示位点名；CHR 表示染色体；BP 表示位点位置；BIP 表示该位点与躁郁症单疾病检测 p 值；SCZ 表示该位点与精神分裂症单疾病检测 p 值；MDD 表示该位点与重度抑郁症单疾病检测 p 值。

MGAS 模型将位点根据 RefGene 数据集，定义位点两侧都有 5 kb 的扩展映射到基因中去，如果一个位点位于多个基因的重叠区域则被认为它属于多个基因。数据共有 454920 个位点位于 17647 个基因。MGAS 根据由 1000 Genomes Project 不相关的欧洲祖先样本基因型数据计算得到的基因内位点间的相关性，生成 LD 矩阵。

1.3　数据分析算法：MGAS

假设基因内有 n 个位点，m 个相关表型，传统的 GWAS 分析工具（PLINK[11]、Mach2dat/DSL[12]、SNPtest[13] 等）使用统计上适当的方法（例如根据表型的测量规模进行线性或逻辑回归）测试 m 个表型与 n 个位点的单疾病关联。MGAS 模型将 GWAS 得到的 $m\times n$ 个 p 值按升序排列从而派生出一个基因水平的 p 值 P_{MGAS}：

$$P_{\mathrm{MGAS}}=\min\left(\frac{q_e}{q_{ej}}P_j\right) \tag{1}$$

其中：q_e 表示一个基因内独立有效的 p 值个数，一个基因内的 p 值总数为 $m\times n$ 个，但由于 p 值是相关的，即 p 值由于表型之间的相关性和位点之间的相关性而相关，因此对有效 p 值数进行了校正；q_{ej} 表示前 j 项 p 值中独立有效的 p 值个数，j 从 1 迭代到 $m\times n$；P_j 指按升序排序后的第 j 个 p 值。因此，P_{MGAS} 是与零假设 H_0 相关的最小加权 p 值，即：

H_0：该基因内的 m 个表型与 n 个位点之间没有关联；

H_1：m 个表型中的至少一个与基因中的 n 个位点中至少一项相关联。

用 P_1 表示表型 m_1 与第 n_1 个位点间的关联测试的 p 值，用 P_2 表示表型 m_2 和第 n_2 个位点之间的关联测试的 p 值。P_1 和 P_2 之间的相关性取决于观察到的第 n_1 个位点与第 n_2 个位点之间的相关性 $r_{n_1n_2}$，以及观察到的表型 m_1、m_2 之间的相关性 $r_{m_1m_2}$。通过对 $m\times n$ 个升序 p 值之间的相关矩阵 $\boldsymbol{\Phi}$ 进行特征值分解，可以得出前 j 个 p 值中有效的 p 值数 q_{ej}。p 值间的相关矩阵 $\boldsymbol{\Phi}$ 虽然无法直接得到，但是可以通过位点间的 $n\times n$ 相关性矩阵 $\boldsymbol{\Omega}$ 和表型间的 $m\times m$ 相关性矩阵 $\boldsymbol{\Sigma}$ 模拟估计得到。q_{ej} 可由式（2）计算：

$$q_{ej}=j-\sum_{i=1}^{j}I(\lambda_i)(\lambda_i-1) \tag{2}$$

其中：j 表示前 j 项 p 值个数；λ_i 表示第 i 个特征值；$I(\lambda_i)$ 为指示函数，当 $\lambda_i<1$ 时取 0，$\lambda_i>1$ 时取 1：

$$I(\lambda_i)=\begin{cases}0, & \lambda_i\leqslant 1\\ 1, & \lambda_i>1\end{cases} \tag{3}$$

p 值有效数 q_{ej} 定义为 p 值个数 j 减去所有特征值大于 1 的 λ_i 与 1 的差之和。如果 j 个 p 值都不相关（即统计检验中涉及的 m 个表型与 n 个位点产生 j 个 p 值，则不相关），则所有 j 个特征值都为 1，那么 $q_{ej}=j-0=j$。相反，如果 j 个 p 值都完全相关（即 m 个表型 n 个位点都完全相关），那么第一个特征值等于 j，其余的特征值为 0，$q_{ej}=j-(j-1)=1$（即测试完全相关的表型与完全相关的位点的关联性只会产生一个唯一的信息单位）。实际上，q_{ej} 通常小于 j 但大于 1，因为表型相关性和位点相关性可能取[−1,1]间的值。q_e 可以看作 j 为 $m\times$

n 时 q_{ej} 的特殊情况。在该模型下，q_e 总是不小于 q_{ej}，所以加权的 P_j 总是不小于没有加权的 P_j。

$m\times n$ 个 p 值的相关性矩阵 $\boldsymbol{\Phi}$ 虽不能直接从样本中得到，但可以通过位点间的 $n\times n$ 相关性矩阵 $\boldsymbol{\Omega}$ 与表型间的 $m\times m$ 相关性矩阵 Σ 的数学关系推导。三者的数学关系由 Li 等[14]关联分析研究使用的数据模拟方法得到，即 p 值相关性矩阵 $\boldsymbol{\Phi}$ 可以通过位点相关性矩阵 $\boldsymbol{\Omega}$ 和表型间相关性矩阵 $\boldsymbol{\Sigma}$ 的 Kronecker 积的六阶多项式函数来精确逼近($R^2=0.9950$)：

$$\boldsymbol{\Phi}=f(\boldsymbol{\Sigma}\otimes\boldsymbol{\Omega}=X)\approx 0.3867X^6+0.0021X^5-0.1247X^4-0.0104X^3+0.7276X^2+0.0068X \tag{4}$$

2 结果与讨论

Lee 等[7]关于精神类疾病遗传相关性的研究，利用混合线性模型处理精神病基因组学联盟提供的全基因组范围的基因型数据，得到三种精神类疾病之间的遗传相关性，其中：躁郁症与精神分裂症的遗传相关性为 0.6801；精神分裂症与重度抑郁症的相关性为 0.4301；重度抑郁症与躁郁症的关联性为 0.4701。此外 Bulik-Sullivan 等[8]通过 LD Hub 模型获取的躁郁症与精神分裂症的遗传相关性为 0.7941；精神分裂症与重度抑郁症的相关性为 0.4781；重度抑郁症与躁郁症的关联性为 0.5083。

为了比较使用两种遗传相关性矩阵多元关联分析结果的差异程度，控制除遗传相关性外的所有变量进行两次独立检验。首先将上述混合线性模型[7]得到的疾病相关性数据作为遗传相关性矩阵，使用原始错误发现率将全基因组范围的显著性阈值调整为 $P=6.8\times10^{-4}$，MGAS 模型检测出 241 个显著性基因如图 1 所示。

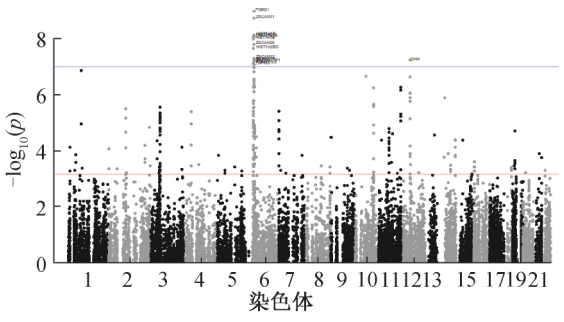

图 1 混合线性模型遗传相关性的多元关联分析结果曼哈顿图

图 1 中发现的基因根据基因位置被映射到染色体上，这里横坐标对应 22 个常染色体，纵坐标表示多元关联分析结果 p 值的负对数值。图中两条阈值线分别为全基因组范围的显著性阈值 $P=6.8\times10^{-4}$ 和基因显著性较高的顶部区域阈值 $P=10^{-7}$，总计发现 241 个显著性基因，15 个顶部区域基因。检测结果相比于 Sklar 等[15]躁郁症研究的原始报告多发现了 100 个显著性基因，相比于 Wray 等[16]重度抑郁症研究的原始报告多发现了 88 个显著性基因，相比于精神分裂症原始报告多发现了 133 个显著性基因。

根据已发布的全基因组关联研究目录(https://www.ebi.ac.uk/gwas/docs/file-downloads)，本文检测出的 241 个显著性基因在关于精神分裂症、躁郁症、重度抑郁症的独立研究中存在一种或多种精神类疾病的相关报道，如在已发布的全基因组关联研究目录中，基因 *PGBD1*($P_{MGAS}=1.04\times10^{-9}$)与精神分裂症具有关联性；基因 *ZSCAN31*($P_{MGAS}=1.82\times10^{-9}$)与重度抑郁症及精神分裂症都存在相关性；基因 *DHH*($P_{MGAS}=5.89\times10^{-8}$)与躁郁症、重度抑郁症及精神分裂症都存在相关性。该结果可以检验已发现的精神类疾病公共致病基因，为发现新的公共致病基因提供方向，如：基因 *PGBD1* 在先前检验中发现与精神分裂症存在较强的关联性，但不同疾病间存在一定的相关性，那么该基因可能为潜在的公共致病基因，在先前重度抑郁症或躁郁症临床检验中因效应量小等原因没有检测出。

观察 15 个顶部区域基因的位点的单疾病 p 值，只有基因 *DHH*($P=5.89\times10^{-8}$)与躁郁症显著相关($P<5\times10^{-8}$)，有 6 个基因只与精神分裂症显著相关，有 7 个基因与精神分裂症和重度抑郁症存在关联。在这 15 个基因中，5 个基因(*PGBD1*、*ZSCAN26*、*ZKSCAN4*、*NKAPL*、*ZNF165*)包含多个与重度抑郁症相关的显著性位点，且在全基因组关联研究目录关于重度抑郁症的研究中没有报告，由于基因内位点过多，结果中显著性较高的部分位点及相关报告数据如表 3 所示。在表 3 中，如果一些位点同时位于多个基因中，则认为该位点同时属于多个基因，如 rs13197574 同时属于 *ZSCAN12P1*、*POM121L2*、*ZNF165*。表 3 中基因 *PGBD1* 的多个位点与重度抑郁症有显著的关联性，但在 GWAS 原始目录中未有相关报告。

类似的，利用 LD Hub 模型提供的遗传相关性再次对三种疾病进行检验，使用原始的错误发现率将全基因组范围的显著性阈值调整为 $P=6.60\times10^{-4}$，MGAS 模型检测出 242 个显著性基因。这里除了包含之前利用混合线性模型提供的疾病遗传相关性检验的 241 个显著性基因外，还包含一个新发现的基因 *RND1*，该基因的 p 值为 $P_{MGAS}=0.0476$。基因 *RND1*

表 3　15 个基因中显著性位点及该基因与疾病已有关联分析报告

Gene	P	P_{MGAS}	Chr	SNP	Feature	BIP	SCZ	MDD	Re
				rs13211507	intronic	3.26×10^{-3}	7.05×10^{-9}	5.80×10^{-9}	SCZ
PGBD1	1.04×10^{-9}	1.61×10^{-5}	6	rs6901575	intronic	5.16×10^{-3}	1.24×10^{-9}	3.40×10^{-9}	SCZ
				rs1936365	intronic	6.40×10^{-3}	6.60×10^{-9}	3.86×10^{-11}	SCZ
ZSCAN31	1.82×10^{-9}	1.61×10^{-5}	6	rs853679	intronic	5.19×10^{-3}	2.11×10^{-9}	7.43×10^{-11}	SCZ,MDD
HIST1H1B	$7,38\times10^{-9}$	3.01×10^{-5}	6	rs13199772	downstream	3.56×10^{-3}	7.05×10^{-10}	8.71×10^{-8}	SCZ,MDD
HIST1H2AL	7.99×10^{-9}	3.01×10^{-5}	6	rs13199772	downstream	3.56×10^{-3}	7.05×10^{-10}	8.71×10^{-8}	SCZ
LINC01012	8.52×10^{-9}	3.01×10^{-5}	6	rs17749927	ncRNA	3.34×10^{-3}	1.01×10^{-9}	2.26×10^{-7}	SCZ
HIST1H1E	1.04×10^{-8}	3.06×10^{-5}	6	rs7749823	upstream	0.19	6.01×10^{-9}	1.14×10^{-4}	SCZ
				rs6901575	intronic	5.16×10^{-3}	1.24×10^{-9}	3.04×10^{-9}	SCZ
ZSCAN26	1.60×10^{-8}	4.03×10^{-5}	6	rs1778508	upstream	2.01×10^{-2}	5.03×10^{-9}	7.06×10^{-9}	SCZ
				rs2799077	upstream	2.33×10^{-2}	5.09×10^{-9}	2.86×10^{-9}	SCZ
HIST1H2BD	2.22×10^{-8}	4.90×10^{-5}	6	rs7749823	upstream	0.19	6.01×10^{-9}	1.14×10^{-4}	SCZ
ZSCAN12	4.95×10^{-8}	8.94×10^{-5}	6	rs13213152	ncRNA	2.11×10^{-4}	3.63×10^{-9}	1.24×10^{-8}	SCZ,MDD
				rs1778508	upstream	2.01×10^{-2}	5.03×10^{-9}	7.06×10^{-9}	SCZ
ZKSCAN4	5.61×10^{-8}	8.94×10^{-5}	6	rs10456362	intronic	2.43×10^{-2}	5.96×10^{-9}	9.40×10^{-9}	SCZ
				rs1679709	exonic	2.91×10^{-2}	9.39×10^{-9}	8.90×10^{-9}	SCZ
DHH	5.89×10^{-8}	8.94×10^{-5}	12	rs7296288	downstream	9.39×10^{-9}	0.02	0.47	SCZ,MDD,BIP
				rs1778508	upstream	2.01×10^{-2}	5.03×10^{-9}	7.06×10^{-9}	SCZ
NKAPL	6.17×10^{-8}	8.94×10^{-5}	6	rs1679709	exonic	2.91×10^{-2}	9.39×10^{-9}	8.90×10^{-9}	SCZ
ZSCAN12P1	6.59×10^{-8}	8.94×10^{-5}	6	rs13197574	downstream	7.14×10^{-3}	4.13×10^{-9}	1.53×10^{-8}	SCZ,MDD
POM121L2	7.53×10^{-8}	9.49×10^{-5}	6	rs13197574	downstream	7.14×10^{-3}	4.13×10^{-9}	1.53×10^{-8}	SCZ
ZNF165	8.54×10^{-8}	1.00×10^{-4}	6	rs13197574	downstream	7.14×10^{-3}	4.13×10^{-9}	1.53×10^{-8}	SCZ

注：Gene 表示基因名；P 表示多元关联分析 p 值；P_{MGAS} 表示 FDR 修正后 p 值；Chr 表示染色体；SNP 表示位点名；Feature 表示位点类型；BIP 表示该位点与躁郁症单疾病检测 p 值；SCZ 表示该位点与精神分裂症单疾病检测 p 值；MDD 表示该位点与重度抑郁症单疾病检测 p 值；Re 表示 GWAS 报告中已有与该基因相关的疾病。

在利用混合线性模型提供的疾病遗传相关性检验中 p 值为 $P_{MGAS}=0.0503$，此外该基因在已发布的全基因组关联研究目录中有其与躁郁症相关的报告。

以上两个结果可以说明，在疾病遗传相关性数据小范围浮动的情况下，MGAS 模型对于两种不同的遗传相关性数据的多元关联分析结果是稳定的。因此，使用 LD Hub 模型提供的疾病间的遗传关系数据作为疾病间相关性矩阵从而将研究范围扩展到更多的疾病是可行的。

3　结　语

当疾病间具有遗传相关性时，基于基因的联合关联分析比单疾病位点水平关联分析更能检测出风险位点。比如 Wang 等[17]的基于基因的单疾病关联分析检测出 9 个与躁郁症相关的基因，相比于之前精神病基因组学联盟中位点水平的荟萃分析多检测出了两个基因：*ANK3*、*SYNE1*；而 Prata 等[18]的躁郁症与精神分裂症的联合分析研究相比于基于基因的单疾病关联分析，又新发现了基因 *CACNAIC*、*ODZ4* 以及显著性区域 NEK4-ITIH1-ITIH3-ITIH4。多表型关联分析同时考虑了与该疾病在病理或临床上相关的多个表型，可能有助于发现潜在的致病基因，如本文通过 MGAS 模型进行躁郁症、重度抑郁症、精神分裂症之间的多元关联分析，新发现 5 个与重度抑郁症相关的基因 *PGBD1*、*ZSCAN26*、*ZKSCAN4*、*NKAPL*、*ZNF165*。将检测得到的显著性基因与先前单疾病研究结果进行对比，可以为确定疾病诊断界限、发现潜在的致病基因提供帮助。对本文检测到的 242 个显著性基因进行通路的富集分析以建立多个疾病遗传重叠的生物学关联有助于识别潜在的神经生物学机制。期望这些发现将最终为精神病学提供信息，并产生新的预防和治疗模式。

多元关联分析中复杂疾病遗传相关性的选择是一大难点。线性混合模型计算遗传相关性时需要用到样本原始的基因型、表型数据，理论上来说更加准确，但受原始数据收集、成本问题等因素影响，有准确遗传相关性结果的疾病种类不多。LD Hub 模型则是使用全基因组关联分析结果 p 值通过线性回归模型来派生出的疾病间遗传相关性，因此可以得更多种疾病间的遗传关系。由于统计的方法总会存在误差，LD Hub 模型

所提供的遗传相关性是否能在多元关联分析中表现出很强的稳定性是本文主要研究的问题。

本文两组多元关联分析结果显示,两种遗传性关性的多元关联分析结果具有一致性,使用 LD Hub 模型提供的疾病间的遗传关系数据作为疾病间相关性矩阵,从而将研究范围扩展到更多的疾病是可行的。通过 LD Hub 模型提供的遗传相关性拓展多元关联分析的疾病选择空间,有助于发现新的易感基因和富集通路,为预测各类疾病-基因的相关性分析提供新的思路。

参考文献:

[1] Manolio T A, Collins F S, Cox N J, et al. Finding the missing heritability of complex diseases[J]. Nature, 2009, 461(7265): 747-753.

[2] Cole J, Ball H A, Martin N C, et al. Genetic overlap between measures of hyperactivity/inattention and mood in children and adolescents[J]. Journal of the American Academy of Child& Adolescent Psychiatry, 2009, 48(11): 1094-1101.

[3] 周家蓬, 裴智勇, 陈禹保, 等. 基于高通量测序的全基因组关联研究策略[J]. 遗传, 2014, 36(11): 1099-1111.

[4] Yang Q, Wu H S, Guo C Y, et al. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests[J]. Genetic Epidemiology, 2010, 34(5): 444-454.

[5] O'Reilly P F, Hoggart C J, Pomyen Y, et al. MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS[J]. PLoS One, 2012, 7(5): e34861.

[6] van der Sluis S, Dolan C V, Li J, et al. MGAS: a powerful tool for multivariate gene-based genome-wide association analysis[J]. Bioinformatics, 2015, 31(7): 1007-1015.

[7] Lee S H, Ripke S, Neale B M, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs[J]. Nature Genetics, 2013, 45(9): 984-994.

[8] Bulik-Sullivan B, Finucane H K, Anttila V, et al. An atlas of genetic correlations across human diseases and traits[J]. Nature Genetics, 2015, 47(11): 1236-1241.

[9] Ward L D, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease[J]. Nature Biotechnology, 2012, 30(11):1095-1106.

[10] Gamazon E R, Wheeler H E, Shah K P, et al. A gene-based association method for mapping traits using reference transcriptome data[J]. Nature Genetics, 2015, 47(9): 1091-1098.

[11] Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses[J]. The American Journal of Human Genetics, 2007, 81(3): 559-575.

[12] Aulchenko Y S, Ripke S, Isaacs A, et al. GenABEL: An R library for genome-wide association analysis[J]. Bioinformatics, 2007, 23(10): 1294-1296.

[13] Li Y, Willer C, Sanna S, et al. Genotype imputation[J]. Annual Review of Genomics and Human Genetics, 2009, 10(1): 387-406.

[14] Li M X, Gui H S, Kwan J S H, et al. GATES: A rapid and powerful gene-based association test using extended simes procedure[J]. The American Journal of Human Genetics, 2011, 88(3): 283-293.

[15] Sklar P, Smoller J W, Fan J, et al. Whole-genome association study of bipolar disorder[J]. Molecular Psychiatry, 2008, 13(6): 558-569.

[16] Wray N R, Ripke S, Mattheisen M, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression[J]. Nature Genetics, 2018, 50(5): 668-681.

[17] Wang M H, Huang J F, Liu Y Y, et al. COMBAT: A combined association test for genes using summary statistics[J]. Genetics, 2017, 207(3): 883-891.

[18] Prata D P, Costa-Neves B, Cosme G, et al. Unravelling the genetic basis of schizophrenia and bipolar disorder with GWAS: A systematic review[J]. Journal of Psychiatric Research, 2019, 114: 178-207.

(责任编辑:康 锋)