浙江理工大学学报，2022，47(6)：923-930
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2022.06.016

基于改进的多元关联分析模型的多种精神疾病相关基因识别

赵国英，贺平安
（浙江理工大学理学院，杭州 310018）

摘　要：基于基因的多元关联分析(Multivariate gene-based association analysis，MGAS)模型能有效地识别基因与表型之间的相关性，然而现有MGAS模型在多元关联分析时会剔除很多相关性数值个数小于疾病种类数目的单核苷酸多态性(Single nucleotide polymorphism，SNP)位点。针对多元关联分析中潜在SNP位点缺失的问题，利用数据填充的方法改进了MGAS模型，并将其应用于6类精神疾病的基因与表型相关性的识别。通过对比改进的MGAS模型与原有模型得到的Top显著基因发现，改进的MGAS模型提高了多元关联分析与疾病相关基因的识别能力，有助于发现多种疾病间的潜在的风险基因，为疾病的预防、诊断和治疗研究提供新的工具和思路。

关键词：精神疾病；多元关联分析；MGAS模型；显著基因；数据填充
中图分类号：O29　　**文献标志码**：A　　**文章编号**：1673-3851 (2022) 11-0923-08

Identification of multiple psychiatric disorders correlated genes based on an improved multivariate association analysis model

ZHAO Guoying, *HE Ping'an*
(School of Science, Zhejiang Sci-Tech University, Hangzhou, 310018, China)

Abstract: Multivariate gene-based association analysis (MGAS) model can effectively identify the genotypes-phenotypes association. However, many single nucleotide polymorphism (SNP) loci with the number of correlation values being less than the number of disease were deleted in the multivariate association analysis of the existing MGAS model. An improved MGAS model based on data completion method was proposed to solve the problem of missing potential SNP loci in multivariate association analysis. And it was applied to identify the genotypes-phenotypes association in six psychiatric disorders. The results of the improved MGAS model and the original model showed that the improved MGAS model improved the identification ability of the genotypes-phenotypes association. The work is conducive to discovering the potential risk genes among various diseases and provide new ideas for disease prevention, diagnosis and treatment.

Key words: psychiatric disorders; multivariate association analysis; MGAS model; significant gene; data completion

0　引　言

全球有3.3%～6.7%的人口患有精神疾病[1]。作为一类复杂疾病，精神疾病已成为中国严重的公共卫生问题。该类疾病会损害患者认知、情绪和行为的能力，具有高发、早发、慢性和反复等特征[2]。

收稿日期：2022－05－11　　网络出版日期：2022－09－06
基金项目：国家自然科学基金项目(61772027)
作者简介：赵国英(1994－　)，女，山西山阴人，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

精神疾病之间存在重叠的临床表型导致很难定义明确的临床诊断界限,例如,精神分裂症、双相情感障碍症与重度抑郁症都具有焦虑、情绪波动和抑郁等临床表型。

研究表明,精神疾病具有高度的遗传性[3]。精神疾病的遗传性由多个效应量很小的位点共同作用[4],因此研究精神疾病遗传分子机制的关键在于定位其风险位点。对此,Risch 等[5]提出了全基因组关联研究(Genome-wide association studies,GWAS)方法,该方法通过在全基因组范围内使用基因分型技术,以标记与疾病相关的单核苷酸多态性(Single nucleotide polymorphism,SNP)。GWAS 被普遍应用于识别疾病与多态性位点之间的相关性,已成为研究疾病遗传机制的主流方法[6-7]。随着分子遗传学的快速发展,基于多种疾病的 GWAS 数据的多元关联分析,可以在分子水平有效地识别它们共同的遗传变异。多种多元关联分析方法已经成功运用到识别复杂疾病的潜在遗传变异中,如多表型联合模型(Joint model of multiple phenotypes,Multiphen)[8]、GWAS 的多性状分析(Muti-Trait analysis of GWAS,MTAG)[9]、多种表型的加权分析(A weighted combination of multiple phenotypes,WCmulP)[10]和基于基因的多元关联分析模型(Multivariate gene-based association analysis,MGAS)[11]等。Yao 等[12]使用 GWAS 的多性状分析方法对精神分裂症、双相情感障碍、重度抑郁症、自闭症谱系障碍和注意力缺陷多动障碍进行多元关联分析,发现了 16 个与多种疾病相关的新基因。Meng 等[13]使用 MGAS 模型对阿尔茨海默病皮质下成像表型进行多元关联分析,发现了 4 个新的显著基因。这些多元关联分析模型在识别与一种或少数具有相关性表型的遗传因素时具有强大的识别能力,但当遗传因素影响多数表型时,识别效率就会降低。

MGAS 模型直接使用来自 GWAS 的原始数据,计算多个表型的基因水平的相关性,不需要对数据中的信息重新整合,提高了模型的使用效力;而且在识别相关基因方面,通过模拟真实的临床数据,表现出了较好的统计能力。然而,MGAS 模型必须使用与所有研究对象都具有相关性指标的位点作为分析数据。这样大量与疾病相关的潜在位点必须被剔除,这些位点极可能是多种疾病多元关联分析的有效研究位点,但由于较小的效应量而未被全基因组关联分析识别出来。

本文针对多元关联分析中潜在位点缺失的问题,对 MGAS 模型进行改进,并将其应用于 6 类精神疾病的基因与表型相关性的识别。首先利用数据填充的方法对 MGAS 模型进行改进,填充原始数据中位点的缺失值,增加 MGAS 模型的分析位点。然后将改进模型应用到精神分裂(Schizophrenia,SCZ)、双相情感障碍(Bipolar disorder,BIP)、重度抑郁症(Major depression,MDD)、注意力缺陷多动障碍(Attention deficit hyperactivity disorder,ADHD)、自闭症谱系障碍(Autism spectrum disorder,ASD)和神经性厌食症(Anorexia nervosa,AN)的多元关联分析中,识别出与多种疾病相关的基因,同时利用已有的 MGAS 模型识别这 6 类精神疾病的相关基因。最后比较改进模型与原有模型的结果,以分析本文方法的特点和有效性。本文改进的多元关联模型可提高多种疾病间的潜在的风险基因的识别能力,进而为精神疾病的预防、诊断和治疗研究提供新的工具和思路。

1 疾病数据与遗传相关性

1.1 疾病数据

本文使用 MGAS 模型对精神分裂症、双相情感障碍、重度抑郁症、自闭症谱系障碍、注意力缺陷多动障碍和神经性厌食症的 GWAS 数据进行多元关联分析。6 类精神疾病的 GWAS 原始数据来源于精神疾病基因组学联盟(https://www.med.unc.edu/pgc)。为了保证多态性位点的连锁不平衡以及样本祖先的一致性,所有数据均采用欧洲血统。注意力缺陷多动障碍的 GWAS 原始数据的部分数据见表 1。根据已下载的 6 类精神疾病的 GWAS 的原始数据,对每个数据中的疾病—对照样本和所含的位点数量进行描述,结果见表 2。

在 6 类精神疾病的 GWAS 原始数据中,神经性厌食症的 GWAS 数据中所含的位点数量最多,精神分裂症的 GWAS 数据中所含的位点数量最少。尽管不同的精神疾病在临床检测中的风险位点存在差异,但仍存在大量重叠的位点,这些位点与多种疾病均具有遗传相关性。

1.2 疾病间的遗传相关性

Bulik-Sullivan 等[14]提出 LD 评分回归(LD score regression,LDSC),对来自全基因组关联分析的原始数据进行回归,以衡量不同疾病之间的遗传相关性。研究表明,由 LD Hub[15]获得的疾病间遗传相关性数据具有可靠性[16]。因此本文使用 LD Hub 提供的疾病间的遗传相关性,以生成多元关联分析所需的疾病相关性矩阵。6 类精神疾病间的遗传相关性数据见表 3。

表 1　注意力缺陷多动障碍的 GWAS 的原始数据

Chr	SNP	Pos	A1	A2	INFO	OR	SE	P
1	rs202152658	751343	A	T	0.880	1.02737	0.0226	0.2333
1	rs143225517	751756	T	C	0.880	0.97336	0.0226	0.2332
1	rs3094315	752566	A	G	0.941	0.096281	0.0194	0.0505

注:Chr 表示染色体;SNP 表示位点;Pos 表示位点所在的位置;A1 表示参考等位基因;A2 表示影响等位基因;INFO 表示插补信息分数;OR 表示优势比;SE 表示比值比标准差;P 表示位点与疾病的关联 p 值。

表 2　6 类精神疾病的 GWAS 数据

疾病名称	病例	对照	位点
BIP	7481	9250	2427220
MDD	16823	25632	9533408
SCZ	9394	12462	1252902
AN	3495	10982	10641224
ADHD	19099	34194	8094094
ASD	6197	7377	6440259

表 3　6 类精神疾病间的遗传相关性

疾病名称	ADHD	AN	ASD	BIP	MDD	SCZ
ADHD	1.00	0.27	0.24	0.23	0.06	0.19
AN	0.27	1.00	0.48	0.79	0.01	0.11
ASD	0.24	0.48	1.00	0.51	0.07	0.14
BIP	0.23	0.79	0.51	1.00	0.04	0.19
MDD	0.06	0.01	0.07	0.04	1.00	0.11
SCZ	0.19	0.11	0.14	0.19	0.11	1.00

注:表中每两个疾病之间对应一个遗传相关数值,其取值范围是[−1,1]。

从表 3 中可以看出:双相情感障碍与神经性厌食症之间的遗传相关性是 0.79,在 6 种精神疾病中表现出最强的遗传相关性;重度抑郁症与其他疾病之间的遗传相关性普遍较弱,其中重度抑郁症与神经性厌食症之间的遗传相关性最弱,仅有 0.01。

2　MGAS 模型及其改进

2.1　MGAS 模型

MGAS 模型[11]对全基因组关联分析中多个 SNP 关联的 p 值信息,利用多元关联分析得到具有多个 SNP 位点的基因与相关疾病之间的相关性 p 值。假设基因中具有和 n 个 SNP,m 个与 SNP 相关的疾病。在零假设条件下,H_0 表示该基因内的和 n 个 SNP 与 m 个疾病无相关性。通过对全基因组关联分析中的关联 p 值进行排序和加权,可以得到基于基因的多元关联 p 值 P_{MGAS},其计算公式如下:

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

其中:参数 q_e 是一个基因中独立有效 p 值的个数。理论上,一个基因的 p 值总数为 $m\times n$,但实际上 p 值的独立有效个数需要修正。因为它不仅取决于疾病之间的遗传相关性,还取决于 SNP 之间的相关性以及疾病与 SNP 之间的相关性。P_j 是 p 值倒序排列之后的第 j 个 p 值,q_{ej} 是前 j 个 p 值中独立有效 p 值的个数,其中 j 的取值范围是[1,$m\times n$]。q_{ej} 的计算公式如下:

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

其中:j 表示位列前面的 p 值数量,λ_i 是第 i 个特征值;$I(\lambda_i)$是指示函数,其公式如下:

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

q_{ej} 可以通过对 $m\times n$ 个倒序排列的 p 值之间的相关矩阵 $\boldsymbol{\Phi}$ 的特征值分解得到。虽然 $m\times n$ 个 p 值之间的相关矩阵 $\boldsymbol{\Phi}$ 不能直接得到,但是可以通过 SNP 之间的相关矩阵 $\boldsymbol{\Omega}$ 和表型之间的相关矩阵 $\boldsymbol{\Sigma}$ 近似得到。p 值的相关矩阵可以通过 SNP 之间的相关矩阵 $\boldsymbol{\Omega}$ 和表型之间的相关矩阵 $\boldsymbol{\Sigma}$ 的 Kronecker 乘积来精确逼近,其数学模型如下:

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

2.2　改进的 MGAS 模型

通常一个 SNP 不仅与一种疾病相关,很可能是多种疾病之间共享的风险位点。由于精神疾病是多基因疾病,单个 SNP 的效应量较小,因此需要对大量的 SNP 数据进行研究。在进行多元关联分析之前,需要处理多种疾病的 GWAS 原始数据,得到完整的多种疾病相关 SNP 位点数据。原有的 MGAS 模型仅使用与所有研究的疾病都相关的公共 SNP 位点的 p 值进行计算。例如,计算表 4 中基因 *MRI6079* 与 6 类精神疾病的相关性时,只有第一个位点(rs3011217)的 p 值被采用。相应的其余 11 个位点的 p 值,因为只含有部分疾病的相关性 p 值而被剔除。显然这种计算方法得到的结果不足以代表基因 *MRI6079* 中所有的位点信息。

表 4 基因 *MRI6097* 中的位点与 6 类精神疾病的相关性

Gene	SNP	Chr	Pos	BIP	MDD	SCZ	ASD	ADHD	AN
	rs3011217	1	44303266	2.53×10^{-1}	5.59×10^{-2}	9.96×10^{-1}	6.84×10^{-3}	1.51×10^{-8}	7.67×10^{-1}
	rs3011220	1	44308867	2.79×10^{-1}	4.50×10^{-2}	NA	5.58×10^{-3}	1.04×10^{-8}	7.24×10^{-1}
	rs7542434	1	44308887	4.65×10^{-1}	2.52×10^{-1}	NA	7.73×10^{-3}	1.52×10^{-4}	5.32×10^{-1}
	rs246774	1	44299695	NA	2.55×10^{-2}	NA	4.97×10^{-1}	8.95×10^{-4}	8.49×10^{-1}
	rs12046114	1	44301820	NA	2.37×10^{-1}	NA	7.69×10^{-3}	1.49×10^{-4}	5.31×10^{-1}
MRI6079	rs246773	1	44300249	NA	2.79×10^{-2}	NA	5.36×10^{-1}	5.81×10^{-4}	9.15×10^{-1}
	rs3791075	1	44300526	NA	2.57×10^{-1}	NA	6.93×10^{-3}	1.48×10^{-4}	5.31×10^{-1}
	rs3838470	1	44303025	NA	2.05×10^{-2}	NA	NA	5.25×10^{-4}	7.90×10^{-1}
	rs3011218	1	44303329	NA	8.20×10^{-2}	NA	6.40×10^{-1}	5.11×10^{-4}	7.91×10^{-1}
	rs3030219	1	44306120	NA	4.73×10^{-2}	NA	5.91×10^{-3}	1.29×10^{-8}	7.71×10^{-1}
	rs2906471	1	44306422	NA	4.72×10^{-2}	NA	5.92×10^{-3}	1.37×10^{-8}	7.70×10^{-1}
	rs2906470	1	44306559	NA	2.76×10^{-2}	NA	5.12×10^{-1}	5.80×10^{-4}	9.27×10^{-1}

注:Gene 表示基因,BIP 表示位点和双相情感障碍的关联 p 值;MDD 表示位点和重度抑郁症的关联 p 值;SCZ 表示位点和精神分裂症的关联 p 值;ASD 表示位点和自闭症谱系障碍的关联 p 值;ADHD 表示位点和注意力缺陷多动障碍的关联 p 值;AN 表示位点和神经性厌食症的关联 p 值;NA 表示位点与疾病之间存在空缺的关联 p 值。

由于现有基因分型技术的限制,以及不同疾病的病例—对照样本在选取时存在实际差异,全基因组关联分析中仍有大量遗传变异尚未被发现。为了解决这个问题,本文应用数据填充的方法将原始数据矩阵的空缺进行填充。因为空缺位点表示现有的 GWAS 分析还没有发现这些位点与相关疾病之间存在相关性,所以本文将所有原始数据矩阵中的空缺值用 1 填充。根据 MGAS 模型的原理,将空缺的关联 p 值填充为 1,既不会影响原始数据中的显著位点,又将所有位点都纳入计算的范围,使得多元关联分析的结果更加全面。本文改进的 MGAS 模型使用欧洲样本的千人基因组计划基因组数据(http://www.1000genomes.org/)计算 SNP 位点间的相关性。本文设置延伸基因区域为 5′端扩展 5 kb 的长度。当一个位点位于多个基因的重叠区域时,该位点将被分配给所有的相关基因。

3 多元关联分析结果及对比

3.1 多元关联分析结果

基于常见疾病/常见变异(CD/CV)假设[17],针对常染色体上等位基因频率大于 0.01 的遗传变异进行研究。利用 MAGS 模型,根据 SNP 与疾病、SNP 之间以及疾病之间的遗传相关性,计算 6 类精神疾病与基因之间的相关性。

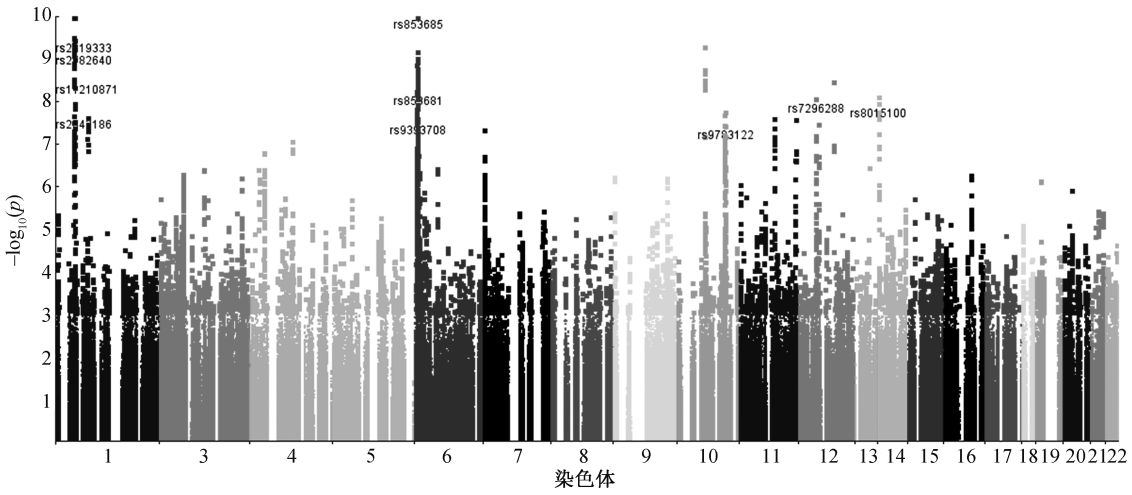
为了快速计算疾病与基因之间的相关性 p 值,MGAS 模型采用了分而治之的算法[11]。对于基因内 $m\times n$ 个 p 值,根据其预期的相关性以 0.05 为步长分为 k 个块。每个块内应用 MGAS 的计算原理确定关键 SNP,对每一个块内的关键 SNP 再次计算获得基因与疾病间的相关性 p 值。

首先,基于原始数据和填充数据,利用 MAGS 模型计算每个 SNP 位点与 6 类疾病之间的相关性,计算结果如图 1 所示。

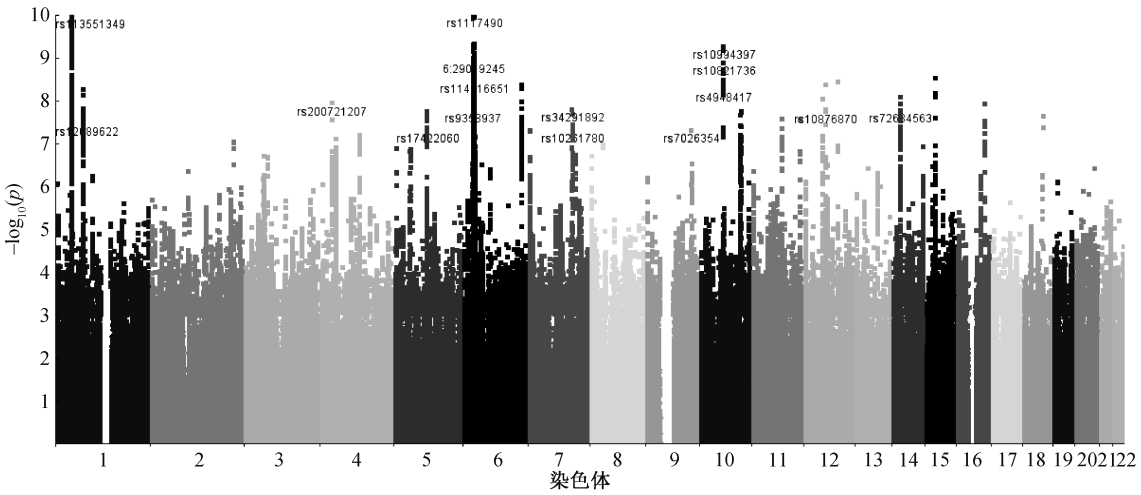
观察图 1 发现,利用 MGAS 模型识别的两个多元关联分析中的关键位点并不相同。只利用原始数据识别的显著位点集中分布在 1 号染色体和 6 号染色体上(图 1(a)),而利用填充数据识别的关键位点集中分布在 1 号染色体、6 号染色体和 10 号染色体上(图 1(b))。其主要原因是填充数据后,所有的 SNP 位点与疾病之间的相关性都被识别。

然后,基于图 1 中的相关性位点,利用 MGAS 模型识别每个基因与疾病的相关性。每个基因与疾病之间的相关性 p 值都有两个 P_{norminal} 和 $P_{\text{corrected}}$。$P_{\text{norminal}}<10^{-7}$ 的基因被定义为 Top 显著基因,$P_{\text{corrected}}<0.05$ 的基因被定义为显著基因[11]。两类模型识别出的 Top 显著基因结果见表 5 和表 6。

利用原始数据,MGAS 模型识别了 303 个显著基因,其中 17 个为 Top 显著基因(表 5)。Top 显著基因有 10 个蛋白质编码基因,分别是:*ST3GAL3*、*KDM4A*、*PGBD1*、*ZSCAN31*、*PTPRF*、*ZSCAN31*、*DUSP6*、*DHH*、*ZKSCAN4* 和 *ZSCAZ1*。利用改进的 MGAS 模型识别了 374 个显著基因,其中 16 个是 Top 显著基因(表 6)。Top 显著基因有 9 个蛋白质编码基因,分别是:*KDM4A*、*ST3GAL3*、*PTPRF*、*PGBD1*、*TRIM26*、*ZSCAN31*、*SLC6A9*、*MDC1* 和 *TUBB*。



(a) 原始数据



(b) 填充数据

图 1　原始数据和填充数据经 MGAS 模型识别的关键位点

注:图中的横坐标是 22 对染色体,纵坐标是与 6 类疾病相关的位点的关联 p 值的负对数。

表 5　原始数据经 MGAS 模型识别的 Top 显著基因

Gene	P_{nominal}	$P_{\text{corrected}}$	Chr	Start_Pos	Group	Length
LOC101929592	1.83×10^{-11}	3.02×10^{-7}	1	44175075	unknown	17939
ST3GAL3	1.52×10^{-10}	8.82×10^{-7}	1	44173206	protein-coding gene	223625
KDM4A	1.60×10^{-10}	8.82×10^{-7}	1	44115819	protein-coding gene	55370
PGBD1	1.60×10^{-9}	6.61×10^{-6}	6	28249313	protein-coding gene	21013
ZSCAN31	2.87×10^{-9}	7.91×10^{-6}	6	28292514	protein-coding gene	11397
PTPRF	2.87×10^{-9}	7.91×10^{-6}	1	43996553	protein-coding gene	92783
H2AC16	9.40×10^{-9}	2.22×10^{-5}	6	27833094	unknown	482
H1-5	1.23×10^{-8}	2.53×10^{-5}	6	27834569	unknown	797
LINC01012	1.92×10^{-8}	3.36×10^{-5}	6	27661857	non-coding RNA	15079
ZSCAN26	2.04×10^{-8}	3.36×10^{-5}	6	28235136	protein-coding gene	10864
H1-4	2.34×10^{-8}	3.51×10^{-5}	6	26156556	unknown	787
H2BC5	3.64×10^{-8}	5.01×10^{-5}	6	26158349	unknown	486
DUSP6	5.74×10^{-8}	7.29×10^{-5}	12	89741013	protein-coding gene	5265
DHH	6.99×10^{-8}	8.25×10^{-5}	12	49480438	protein-coding gene	8146
ZKSCAN4	7.84×10^{-8}	8.38×10^{-5}	6	28209474	protein-coding gene	10573
MIR6079	8.12×10^{-8}	8.38×10^{-5}	1	44304293	non-coding RNA	62
ZSCAN12	8.95×10^{-8}	8.69×10^{-5}	6	28352514	protein-coding gene	2132

注:P_{nominal} 表示经 MGAS 模型识别的基因的 p 值,$P_{\text{corrected}}$ 表示该基因的矫正 p 值,Start_Pos 表示该基因的起始位置,Group 表示该基因的类型,Length 表示该基因的长度。

表 6 填充数据经 MGAS 模型识别的 Top 显著基因

Gene	P_{norminal}	$P_{\text{corrected}}$	Chr	Start_Pos	Group	Length
KDM4A-AS1	5.48×10^{-11}	9.49×10^{-7}	1	44165408	non-coding RNA	7604
LOC101929592	7.17×10^{-11}	9.49×10^{-7}	1	44175075	unknown	17939
KDM4A	1.59×10^{-10}	1.40×10^{-6}	1	44115819	protein-coding gene	55370
ST3GAL3	6.40×10^{-10}	4.23×10^{-6}	1	44173206	protein-coding gene	223625
PTPRF	3.58×10^{-9}	1.89×10^{-5}	1	43996553	protein-coding gene	92783
PGBD1	4.56×10^{-9}	2.01×10^{-5}	6	28249313	protein-coding gene	21013
TRIM26	1.09×10^{-8}	4.10×10^{-5}	6	30152231	protein-coding gene	28952
ZSCAN31	1.77×10^{-8}	5.86×10^{-5}	6	28292514	protein-coding gene	11397
SLC6A9	2.01×10^{-8}	5.90×10^{-5}	1	44462154	protein-coding gene	20681
H2AC15	3.95×10^{-8}	8.82×10^{-5}	6	27805657	unknown	496
H2BC15	4.15×10^{-8}	8.82×10^{-5}	6	27806346	unknown	542
H2AC16	4.20×10^{-8}	8.82×10^{-5}	6	27833094	unknown	482
H1－5	4.33×10^{-8}	8.82×10^{-5}	6	27834569	unknown	797
MDC1	7.93×10^{-8}	1.43×10^{-4}	6	30667583	protein-coding gene	17475
H1－4	8.13×10^{-8}	1.43×10^{-4}	6	26156556	unknown	787
TUBB	8.74×10^{-8}	1.45×10^{-4}	6	30689414	protein-coding gene	3781

3.2 多元关联分析结果的对比

比较表 5 和表 6 的结果可以发现两个模型的结果存在以下差异：

利用填充数据 MGAS 模型计算得到的基因与疾病关联 p 值高于利用原始数据 MGAS 模型计算的 p 值。例如，基因 *LOC101929592*、*ST3GAL3*、*KDM4A*、*PGBD1*、*ZSCAN31*、*PTPRF*、*H2AC16*、*H1-5*、*LINC01012*、*ZSCAN26*、*H1-4*、*H2BC5*、*DUSP6*、*DHH*、*ZKSCAN4*、*MIR6079* 和 *ZSCAN12*。这是由于改进的 MGAS 模型中新增的关联 p 值为 1 的位点会影响基因水平的 p 值，其中基因 *LOC101929592*、*ST3GAL3*、*KDM4A*、*PGBD1*、*ZSCAN31*、*PTPRF*、*H2AC16*、*H1-5* 和 *H1-4* 利用填充数据 MGAS 模型计算的 p 值仍然小于 1×10^{-7}。基因 *LINC01012*、*ZSCAN26*、*H2BC5*、*DUSP6*、*DHH*、*ZKSCAN4*、*MIR6079* 和 *ZSCAN12* 利用填充数据 MGAS 模型计算的 p 值大于 1×10^{-7}。

利用填充数据 MGAS 模型计算得到的基因与疾病关联 p 值低于利用原始数据 MGAS 模型计算的 p 值。例如基因 *KDM4A-AS1*、*SLC6A9*、*H2AC15* 和 *H2BC15*。利用原始数据的 MGAS 模型计算得到基因 *H2AC15* 和 *H2BC15* 显著基因不是 Top 显著基因，基因 *KDM4A-AS1* 和 *SLC6A9* 甚至不是显著基因。产生这种情况的原因是：基因内有效 SNP 数量增加导致之前独立的 SNP 不再独立。根据 MGAS 模型的原理，基因内独立有效 SNP 个数发生改变，其 p 值会随之改变。例如：基因 *KDM4A-AS1* 在原始数据中只包含 rs3791039 和 rs7528454 这 2 个位点，且它们与 6 类精神疾病的关联 p 值都大于 1×10^{-4}，然而在改进模型计算中增加了 7 个位点，其中位点 rs56319043 和 rs112984125 与注意力缺陷多动障碍的关联 p 值分别是 1.37×10^{-11} 和 1.08×10^{-12}，都低于全基因组显著阈值（5×10^{-8}）[18]。因此基因 *KDM4A-AS1* 由改进的 MGAS 模型获得的多元关联 p 值（$P_{\text{norminal}}=5.48\times10^{-11}$）低于原有模型得到的 p 值（$P_{\text{norminal}}=1.32\times10^{-2}$）。

利用填充数据 MGAS 模型计算的结果中有 3 个原有模型没有的 Top 显著基因，即 *TRIM26*、*MDC1* 和 *TUBB*。这是因为基因 *TRIM26* 包含 105 个位点，*MDC1* 包含 14 个位点，*TUBB* 包含 20 个位点，在原始 GWAS 数据中，这 3 个基因中的所有位点至少与 1 类精神疾病的关联 p 值缺失。因此利用原始数据，MGAS 模型不能识别它们与疾病的相关性。而在改进的模型中对其进行数据填充后作为计算数据，所有的位点数据都被用作了模型的计算数据。这说明，改进的 MGAS 模型可以得到原有模型无法识别出的基因。

搜索 GWAS Catalog（https://www.ebi.ac.uk/gwas/）和 Gene Card（https://www.genecards.org/）数据库发现，基因 *TRIM26*、*MDC1* 和 *TUBB* 与 6 类精神疾病具有不同程度的关联，其中基因 *TRIM26*[19] 与精神分裂症、双相情感障碍、重度抑郁症、注意力缺陷多动障碍和自闭症谱系障碍这 5 类精神疾病相关。基因 *MDC1* 与双相情感障碍、重度抑郁症和自闭症谱系障碍相关。基因 *TUBB* 与精神分裂症、重度抑郁症和自闭症谱系障碍相关。

此外,基因 *TRIM26* 除了 5 类精神疾病相关以外,还与神经管缺陷疾病相关。*TRIM26* 蛋白[20]由一个 RING 结构域、一个 B2 结构域、一个卷曲螺旋结构域和一个 PRY-SPRY 结构域组成。已有研究表明,该基因在呼吸疾病[21]、肿瘤[22]等疾病中发挥重要作用,而且被鉴定为精神分裂症的易感基因[23]。研究表明,通过 Meta 分析鉴定出三个基因 *RNF5*、*HLA-DRB3* 和 *TRIM26* 基因被多个 SNP 调节,这三个基因均位于主要组织相容性复合体(major histocompatibility complex,MHC)区域内,最重要的是其表达与精神分裂症相关。公开可用的大脑表达数据表明,基因 *TRIM26* 和 *HLA-DRB3* 的表达数量性状位点也存在于特定的大脑区域[24]。

基因 *MDC1* 为 kiaa 家族成员之一,其编码的蛋白含有 FHA 结构域、BRCT 结构域、PST 结构域。细胞周期检查点和蛋白质代谢是该基因的相关通路。*MDC1* 蛋白与组蛋白 H2A 家族的成员之一 H2AX 相互作用,促进 ATM 激酶和减数分裂重组 11 蛋白复合物向 DNA 损伤灶的募集。

基因 *TUBB* 与 α—微管蛋白相互作用形成二聚体,成为微管的结构成分。富含 *TUBB3*[25]的微管比其他 β—微管蛋白组成的微管更具活力,并且它的表达主要限于神经元。*TUBB3* 的转录水平在成人中枢神经系统中较低,而在周围神经系统中仍保持高水平的表达。因此,*TUBB3* 能对神经系统发育和轴突维持具有特定的功能,它的突变可能导致皮质发育不良,并具有其他脑畸形的症状。

这些结果表明,改进的 MGAS 模型解决了多元关联分析中潜在位点缺失的问题,能更有效准确地识别与多种精神疾病相关的基因。

4　结　语

目前的精神疾病诊断统计手册中对疾病的诊断界限并没有提供遗传学的证据,精神疾病具有表型重叠而难以明确诊断界限的问题。在影响精神疾病的众多因素中,遗传因素占据主要地位。GWAS 作为研究单个疾病的遗传物质的有力工具,被广泛应用于研究患病个体基因型与疾病表型之间的遗传相关性。随着分子遗传学的发展,基于 GWAS 的多元关联分析可以有效识别多种疾病的潜在的共同致病基因。

MGAS 模型可以直接使用临床数据模拟基因型与表型之间的关系,具有较高的统计能力。但在进行多元关联分析时,MGAS 模型会剔除关联数值个数与疾病种类数目不一致的位点,导致分析数据中的位点信息不足以涵盖基因内的全部位点信息。针对多元关联分析中潜在位点缺失的问题,本文改进了 MGAS 模型,对空缺的 p 值进行数据填充,避免遗漏因效应量较小而未被识别出的潜在遗传风险。将改进的模型应用于 6 类精神疾病的多元关联分析中,通过比较改进的 MGAS 模型与原有模型得到的 Top 显著基因发现,改进的 MGAS 模型解决了多元关联分析中潜在位点缺失的问题,能更好地识别多种疾病间潜在的风险基因,为精神类疾病的预防、诊断和治疗研究提供新的工具和思路。

参考文献:

[1] Caves Sivaraman J J, Naumann R B. Estimating the association between mental health disorders and suicide: A review of common sources of bias and challenges and opportunities for US-based research [J]. Current Epidemiology Reports, 2020, 7(4): 352-362.

[2] Whiteford H A, Ferrari A J, Degenhardt L, et al. The global burden of mental, neurological and substance use disorders: An analysis from The Global Burden of Disease Study 2010 [J]. PLoS One, 2015, 10(2): e0116820.

[3] Marian A J. Molecular genetic studies of complex phenotypes[J]. Translational Research, 2012, 159(2): 64-79.

[4] Gallagher M D, Chen-Plotkin A S. The post-GWAS era: From association to function[J]. American Journal of Human Genetics, 2018, 102(5): 717-730.

[5] Risch N, Merikangas K. The future of genetic studies of complex human diseases[J]. Science, 1996, 273(5281): 1516-1517.

[6] Fabbri C, Serretti A. Role of 108 schizophrenia-associated loci in modulating psychopathological dimensions in schizophrenia and bipolar disorder [J]. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2017, 174(7): 757-764.

[7] Grove J, Ripke S, Als T D, et al. Identification of common genetic risk variants for autism spectrum disorder[J]. Nature Genetics, 2019, 51(3): 431-444.

[8] 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.

[9] Turley P, Walters R K, Maghzian O, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG[J]. Nature Genetics, 2018, 50(2): 229-

237.

[10] Zhu H H, Zhang S L, Sha Q Y. A novel method to test associations between a weighted combination of phenotypes and genetic variants[J]. PLoS One, 2018, 13(1): e0190788.

[11] 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, 2014, 31(7): 1007-1015.

[12] Yao X M, Glessner J T, Li J Y, et al. Integrative analysis of genome-wide association studies identifies novel loci associated with neuropsychiatric disorders [J]. Translational Psychiatry, 2021, 11: 69.

[13] Meng X L, Li J, Zhang Q S, et al. Multivariate genome wide association and network analysis of subcortical imaging phenotypes in Alzheimer′s disease [J]. BMC Genomics, 2020, 21(S11): 896.

[14] Bulik-Sullivan B K, Loh P R, Finucane H K, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies [J]. Nature Genetics, 2015, 47(3): 291-295.

[15] Zheng J, Erzurumluoglu A M, Elsworth B L, et al. LD Hub: A centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis [J]. Bioinformatics, 2016, 33(2): 272-279.

[16] 侯龙傲,贺平安. 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[J]. 浙江理工大学学报(自然科学版), 2020, 43(5): 687-692.

[17] Reich D E, Lander E S. On the allelic spectrum of human disease[J]. Trends in Genetics, 2001, 17(9): 502-510.

[18] Hindorff L A, Sethupathy P, Junkins H A, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits[J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(23): 9362-9367.

[19] Lontan A, Fenckova M, Bralten J, et al. Neuroinformatic analyses of common and distinct genetic components associated with major neuropsychiatric disorders [J]. Frontiers in Neuroscience, 2014, 8: 331-346.

[20] 廉梅,林晓琳,贾俊双,等. TRIM26 在人和小鼠主要组织器官中的表达谱[J]. 中国比较医学杂志, 2020, 30(8): 10-15.

[21] Pividori M, Schoettler N, Nicolae D L, et al. Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: Genome-wide and transcriptome-wide studies [J]. The Lancet Respiratory Medicine, 2019, 7(6): 509-522.

[22] Sawai H, Nishida N, Khor S S, et al. Genome-wide association study identified new susceptible genetic variants in HLA class I region for hepatitis B virus-related hepatocellular carcinoma[J]. Scientific Reports, 2018, 8: 7958.

[23] Irish Schizophrenia Genomics Consortium and the Wellcome Trust Case Control Consortium 2. Genome-wide association study implicates HLA-C * 01:02 as a risk factor at the major histocompatibility complex locus in schizophrenia[J]. Biological Psychiatry, 2012, 72(8): 620-628.

[24] De Jong S, Van Eijk K R, Zeegers D W L H, et al. Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes[J]. European Journal of Human Genetics, 2012, 20(9): 1004-1008.

[25] Whitman M C, Barry B J, Robson C D, et al. $TUBB_3$ Arg262His causes a recognizable syndrome including CFEOM3, facial palsy, joint contractures, and early-onset peripheral neuropathy [J]. Human Genetics, 2021, 140(12): 1709-1731.

(责任编辑:康 锋)