带有缺失基因型观测的家族数据全基因组关联研究*

北京大学公共卫生学院生物统计系(100191) 王敬元 张翌玺 宋倩倩 梁宝生△

【提 要】 **目的** 利用带有缺失基因型观测的家族关联数据(correlated family data, CFD)和全基因组关联研究(genome-wide association study, GWAS)探索阿尔茨海默病的潜在致病基因及关联强弱。**方法** 研究人群来自华盛顿高地-英伍德哥伦比亚老龄化项目(Washington Heights-Inwood Columbia Aging Project, WHICAP),该项目收集了先证者的基因型信息,并通过调查访谈收集了先证者及其家庭成员的人口统计学信息。本研究纳入 352 名先证者和 820 名关联家庭成员共 1172 人,首先利用家族结构信息和孟德尔遗传定律估计家庭成员缺失基因型的概率分布,然后应用混合效应 logistic 回归模型,并利用极大似然估计和 EM 算法估计基因效应值。最后,将分析结果分别与仅用先证者信息的 logistic 回归模型和使用主成分校正的模型进行对比。**结果** 该 GWAS+CFD 研究新发现了 7 个显著的单核苷酸多态性(single nucleotide polymorphisms, SNP)位点,其中 4 个 SNPs 对应已知的基因位点,分别为 *rs*7918428(DNAJC12, $OR=2.362$, $P=1.82\times10^{-9}$), *rs*6135509(MACROD2, $OR=2.238$, $P=7.40\times10^{-9}$), *rs*4750496(FRMD4A, $OR=2.454$, $P=1.12\times10^{-8}$), rs4721323(MAD1L1, $OR=1.593$, $P=2.04\times10^{-8}$),另外 3 个为新发现的 SNP 位点:*rs*764009($OR=2.321$, $P=2.23\times10^{-8}$), *rs*7593443($OR=1.745$, $P=2.83\times10^{-8}$)和 *rs*2170560($OR=2.603$, $P=3.11\times10^{-8}$)。相比于其他方法,本研究提出的方法能获得最小的基因膨胀系数($\lambda=1.007$)。**结论** 本研究提出了一种新型混合效应 logistic 回归模型来进行带有缺失基因型观测的家族关联数据全基因组关联分析,通过加入多层随机效应有效控制了混杂因素,能显著提高检出遗传变异的能力。该方法发现了多个阿尔茨海默病的潜在致病风险位点,将有助于后续的疾病通路探索,也为疾病检测和治疗药物的研发提供了更多可能性。

【关键词】 阿尔茨海默病 家系研究 全基因组关联研究 缺失数据
【中图分类号】 R195.1 【文献标识码】 A **DOI** 10.11783/j.issn.1002-3674.2024.05.015

阿尔茨海默病(Alzheimer's disease, AD)是一种与年龄有关的以认知功能衰退为表现的神经精神性疾病[1],多发于老年人,且发病率随年龄增加呈快速上升趋势。以我国为例,截至 2018 年,中国 60 岁以上 AD 患病者约 983 万人,居世界首位,预计到 2050 年将超过 4000 万,且因 AD 死亡人数将超过 47 万[2]。近年来,我国老龄化趋势不断加深,老年人健康问题是当前社会的紧迫问题。2021 年发布的《中国阿尔茨海默病患者诊疗现状调研报告》[3]指出,我国 AD 防治仍面临着疾病认知不足、患病率高、就诊率低等困境,而提高全社会的 AD 预防意识,倡导早筛、早诊、早治,是预防 AD 的关键。AD 有明显的遗传倾向,研究显示一级亲属患病可以使 AD 发病风险增加约两倍[4]。因此,研究 AD 相关的遗传因素对疾病诊断和治疗策略制定至关重要。

AD 受到遗传因素和环境因素的共同作用,全基因组关联研究(genome-wide association study, GWAS)是研究复杂疾病相关遗传因素的有效方法。现有的 AD 全基因组关联研究已经发现包括 *ABCA*7, *APOE*-ε4 等在内的 20 多个基因位点[5-7]。对基因风险和遗传通路的研究显示,可能存在罕见变异与 AD 的发病密切相关[7]。最近一项关于英国生物银行(UK Biobank, UKB)AD 数据的 GWAS 研究以父母的表型作为参与者的代理表型,并结合家族史数据,发现了 4 个新的基因位点,其中 *TOMM*40 基因中的单核苷酸多态性(single nucleotide polymorphisms, SNP)与 AD 的关联可能通过前额皮层的基因表达和 DNA 甲基化实现[8]。可见,利用家族数据有利于发现与疾病相关的危险基因。但是在实际研究中,虽然先证者及其家庭成员的表型信息容易获得,但由于遗传信息采集的高昂成本及年长亲属死亡等原因,通常仅有先证者的基因信息可以被收集,从而带来了带有缺失基因型观测的家族数据。而现有的 GWAS 方法无法处理仅有亲属表型而无基因型的家族关联数据(correlated family data, CFD)。

因此,本文提出了一种混合效应 logistic 回归模型,用于亲属基因型缺失的家族关联数据全基因组关联研究。该模型纳入了家庭间随机效应和个体随机效应,既可以很好地刻画家庭成员共有的特征,也可以刻画家庭成员之间的个体差异,从而有效地控制人口学混杂,提高检出遗传变异的能力。本文使用该模型分析了华盛顿高地-英伍德哥伦比亚老龄化项目的数据,旨在利用家族关联数据探索 AD 的遗传风险,揭示其潜在的遗传通路,为 AD 药物研发和治疗提供新的思路。

*基金项目:国家自然科学基金青年项目(11901013);北京市自然科学基金青年项目(1204031);中央高校基础研究基金(BMU2021RCZX023);北京大学人民医院研究与发展基金临床医学+X 培育项目(RDX2021-05)
△通信作者:梁宝生,E-mail:liangbs@hsc.pku.edu.cn

资料与方法

1. 研究对象

本研究在家族关联数据中采用全基因组关联研究(GWAS+CFD),以先证者的基因型及其亲属的表型为研究对象,样本来自华盛顿高地-英伍德哥伦比亚老龄化项目(Washington Heights-Inwood Columbia Aging Project, WHICAP)。该项目是由美国国家老龄化研究所和国家人类基因组研究所开发的以社区为基础、以老年群体衰老和痴呆为研究对象的大型综合性纵向研究,研究对象来自曼哈顿北部社区[9],研究的先证者是从当地医疗保险记录中确定的65岁及以上的老年人。该研究于1992年和1999年招募研究对象,并对其中三个种群(白种人、西班牙裔、非裔美国人)的调查对象进行结构化访谈,收集其亲属的表型;在研究期间,密切追踪随访先证者的AD和痴呆发病情况。由于西班牙裔老年群体AD的发病率约为白人的两倍[10],所以我们选择西班牙裔群体为最终研究对象。另外,考虑到家系成员中儿童和青少年成员通常还没到AD的发病年龄段,故而被排除在分析之外。因此,本研究考虑的家族谱系包括先证者及其父母和兄弟姐妹。本研究所用的基因测序数据是在美国国家阿尔茨海默病衰老遗传研究所申请(编号2021KT63),相关数据在该遗传研究所的数据存储网站(https://www.niagads.org/adsp/content/home)下载。

2. 基因数据和质量控制

本研究纳入的基因位点均经过质量控制,SNPs位点的排除标准为:①基因型缺失率> 5%;②次等位基因频率(minor allele frequency, MAF)<5%;③父母偏离Hardy-Weinberg平衡状态($P<1\times10^{-6}$),筛选通过PLINK(v1.9)软件完成。最终本研究纳入了565170个SNPs。

3. 统计学分析

(1)缺失数据处理

在CFD数据中,亲属的基因型是未知的。定义G_{ij}为第i个家庭中第j个成员一个单核苷酸多态性的基因型,$j=0$表示此人为该家庭中的先证者。在建模分析中,G_{ij}被编码为0、1或2(举例:不妨设该基因为显性基因,则AA=2, Aa=1, aa=0)。数据中,仅先证者群体能观测到基因型,即G_{i0}已知,而先证者的家庭成员基因型均缺失。在给定先证者基因型G_{i0}的条件下,利用完整家系结构和MAF信息,可以估计家庭成员基因型的概率分布。因此,在孟德尔遗传定律下结合蒙特卡洛方法和计算机数值模拟,我们首先估计了所有亲属基因型的联合条件概率分布。

(2)混合效应logistic回归模型

根据每一个家庭中先证者的基因型及其父母和兄弟姐妹的表型信息,某个个体的发病风险π_{ij}可以通过以下混合效应logistic回归模型建模:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)=\beta_0+\beta G_{ij}+\gamma TX_{ij}+b_i+r_{ij},$$

$$b_i\sim N(0,\sigma_b^2),\ r_i=(r_{i0},\cdots,r_{in_i})^T\sim MVN(0,\sigma_r^2\Sigma_i).$$

其中X_{ij}表示年龄、受教育年限等人口信息学信息,Σ_i为基于孟德尔遗传定律的家系矩阵(kinship matrix),用来表示家庭成员之间的遗传相似性,$j=0,1,\cdots,n_i$, $i=1,\cdots,n$。为了调整家庭成员之间的相关性以及多基因效应带来的混杂,模型引入了基于谱系结构计算的个体随机效应r_{ij};同时,考虑到家庭内部成员共同生活方式的影响,模型引入了家庭特异的随机效应b_i以刻画家庭异质性。我们假设二者独立并分别服从多元正态分布和正态分布。为了避免EM算法中的高维重积分,降低计算复杂度,当家族谱系兄弟姐妹人数大于2时,我们随机抽取2位组成一个家庭单位,然后多次重复抽取以覆盖全部家庭成员,并将每次抽取形成的家庭当作独立增强样本纳入分析。把G_{ij}和随机效应项b_i, r_{ij}当作缺失数据,使用极大似然估计和EM算法估计特定基因的效应值(*OR*值)。

(3)模型比较

GWAS容易受到包括人口结构分层和隐匿相关性等人群结构因素的混杂干扰。本文提出的方法理论上可以控制人口结构带来的混杂,为检验其性能,我们进行了三种模型的对比。M1:仅用先证者信息的logistic回归;M2:也是仅使用先证者信息,同时采用调整人群结构的常用方法:使用SNPs获得主成分(principal components, PCs)并作为协变量纳入模型[11];M3:本研究提出的结合家族关联数据和混合效应logistics回归的方法。M2和M3均将性别、年龄和受教育程度等可能与AD相关的人口统计学信息纳入模型[12]。我们使用基因膨胀系数λ来量化因人群结构带来的混杂,理想的膨胀系数为1,实际膨胀系数越偏离1,说明越容易有假阳性结果,需要重新校正人群结构的影响。以上三个模型的统计学分析均使用R软件(v4.0.2)完成。

结　果

1. 基本信息

表1总结了所有研究对象的人口学信息,最终被纳入研究的有来自352个家庭的352名先证者和820名家庭成员,共1172人,其中143人患阿尔茨海默病(12.2%)。

表 1 家族结构研究对象的人口学信息

	先证者	家庭成员	全部
总人数	352	820	1172
性别			
男	258(73.3%)	392(47.8%)	650(55.5%)
女	94(26.7%)	428(52.2%)	522(44.5%)
年龄(岁)	76.05±6.46	66.56±19.01	69.41±16.86
受教育年限(年)	11.44±3.60	10.63±3.89	10.88±3.82
AD 患病情况			
患病	109(31.0%)	34(4.2%)	143(12.2%)
不患病	243(69.0%)	786(95.8%)	1029(87.8%)

2.全基因关联分析

我们在混合效应 logistic 回归模型,即 M3 中,通过 GWAS 研究所得的曼哈顿图如图 1 所示,纵坐标表示 SNP 的$-logP$ 值,横坐标表示 SNP 在染色体上的位点,红色横线表示 GWAS 通用的显著性阈值 $\alpha=5\times10^{-8}$,蓝色横线表示显著性阈值 $\alpha=1\times10^{-5}$。详细结果见表 2。M3 得到了 7 个达到统计学显著性水平($P<5\times10^{-8}$)的 SNPs。其中 4 个为已知基因对应的 SNPs 位点,分别为 *rs*7918428(DNAJC12,$OR=2.362$,$P=1.8\times10^{-9}$),*rs*6135509(MACROD2,$OR=2.238$,$P=7.40\times10^{-9}$),*rs*4750496(FRMD4A,$OR=2.454$,$P=1.12\times10^{-8}$),*rs*4721323(MAD1L1,$OR=1.593$,$P=2.04\times10^{-8}$),以及三个新发现的 AD 相关 SNP 位点:*rs*764009($OR=2.321$,$P=2.23\times10^{-8}$),*rs*7593443($OR=1.745$,$P=2.83\times10^{-8}$)和 *rs*2170560($OR=2.603$,$P=3.11\times10^{-8}$)。同时对比发现,上述 7 个 SNP 位点在 M1 和 M2 中均未达到显著性阈值 $\alpha=5\times10^{-8}$。

仅使用先证者信息以及使用先证者信息加主成分校正的方法,即 M1 和 M2,均未得到在 5×10^{-8}下显著的 SNP。M1 和 M2 得到的显著性最强的 SNP 均为 *rs*2082768(对应基因 LNX1),其在 M1 和 M2 分析中的 P 值分别为 1.80×10^{-6}和 8.12×10^{-7}。该 SNP 在 M3 中的 P 值为 7.21×10^{-8},亦不显著。

3.模型比较

经计算,M1 在仅含先证者信息的情况下,基因膨胀系数 $\lambda=1.044$,表示可能存在人群结构因素带来的混杂;M2 在先证者中使用主成分分析法校正,即在模型中加入利用 SNPs 信息计算的前 10 个主成分,此时得到基因膨胀系数 $\lambda=1.033$;M3 使用本文提出的多层随机效应模型进行校正,基因膨胀系数 $\lambda=1.007$。由此可见,我们的方法有效控制了人群结构因素等带来的混杂。

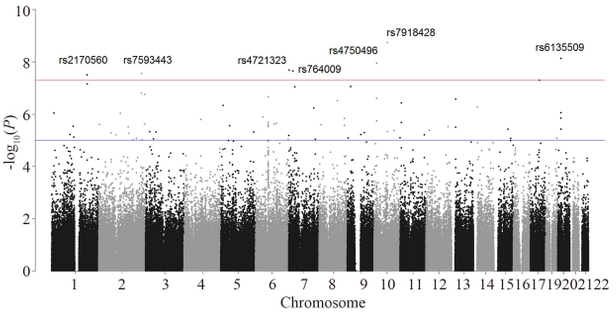

图 1 使用 M3 进行 GWAS 研究的曼哈顿图

表 2 GWAS+CFD 分析得到的阿尔茨海默病相关 SNPs 及相应 SNPs 在单独先证者分析中的结果

Top SNP	染色体数:位置	MA	MAF	M3		M1	
				OR(95%置信区间)	*P* 值	*OR*(95%置信区间)	*P* 值
*rs*7918428	10:69261268	A	0.265	2.362(1.607~3.471)	1.82×10^{-9}	2.014(1.472~2.754)	1.18×10^{-5}
*rs*6135509	20:15707361	G	0.222	2.238(1.598~3.135)	7.40×10^{-9}	2.120(1.537~2.925)	4.69×10^{-6}
*rs*4750496	10:14324612	A	0.371	2.454(1.679~3.586)	1.12×10^{-8}	1.694(1.298~2.210)	1.03×10^{-4}
*rs*4721323	7:1841821	C	0.053	1.593(1.062~2.389)	2.04×10^{-8}	3.493(1.841~6.626)	1.29×10^{-4}
*rs*764009	7:22868317	C	0.255	2.321(1.631~3.301)	2.23×10^{-8}	1.798(1.336~2.418)	1.06×10^{-4}
*rs*7593443	2:221957616	T	0.088	1.745(1.142~2.667)	2.83×10^{-8}	2.436(1.516~3.915)	2.34×10^{-4}
*rs*2170560	1:185303053	G	0.435	2.603(1.675~4.047)	3.11×10^{-8}	1.560(1.204~2.021)	7.58×10^{-4}

* MA:次等位基因(minor allele);MAF:次等位基因频率(minor allele frequency);*OR*:次等位基因拷贝的优势比(odds ratio for per copy of minor allele);M3:本文提出的方法(proposed multilevel GWAS+CFD analysis);M1:仅使用先证者信息进行 GWAS 分析(proband-alone GWAS analysis)

讨 论

在实际研究中,通常仅先证者的基因型能被获得,而亲属的基因型信息往往缺失,传统 GWAS 研究的方法不能解决这一数据缺失问题,给基因与疾病遗传关联的检测带来了挑战。本文提出了一个混合效应 logistic 回归模型,以研究带有先证者全基因组信息和亲属表型信息的家族关联数据中遗传标记与表型之间的关联。该模型在控制了家庭相关性、多基因效应、人群结构因素等混杂的同时,提高了发现遗传变异与表型之间关联的能力。通过对目标数据的实证分析,我们发现了一些潜在的 AD 风险基因,有助于后续 AD 的疾病通路探索,为疾病的检测和治疗提供新的可能性。虽然实证分析使用的是两个兄弟姐妹的谱系,但该方法可以进一步扩展,应用到更大的谱系。

FRMD4A 是已知的 AD 风险基因位点[13]。研究显示 FRMD4A 在功能上可能通过激活细胞粘附素-Arf6 信号传导来调节 tau 分泌,进而影响疾病进

展[14]。本文在西班牙裔的 GWAS + CFD 中发现了 FRMD4A 的统计学显著性（$P = 1.12 \times 10^{-8}$，$OR = 2.45$），这与一项欧洲人群中的全基因组单倍型关联研究结论类似，他们发现 FRMD4A 基因携带者患 AD 的几率增加了 1.43～1.96 倍[13]。而我们的结果进一步表明，先前确认的 FRMD4A 不仅在先证者中是 AD 的易感位点，而且是先证者家庭成员的易感位点。

MACROD2 是人类单 ADP-核糖基化酶的一个编码基因，最近有研究表明 MACROD2 的编码突变与多种神经系统疾病相关，如自闭症、注意力缺陷多动障碍、精神分裂等[15-16]。自闭症的 GWAS 显示 MACROD2 基因异常与自闭症谱系障碍有着显著关联[17-18]。本研究发现 MACROD2 基因（*rs*6135509，$P = 7.40 \times 10^{-9}$）的变异与 AD 高度相关，这可能与 MACROD2 基因影响精神疾病发生的共同通路有关，该发现可以为探索 AD 的疾病通路提供新的线索。

在 GWAS + CFD 中发现的 MAD1L1 基因（*rs*4721323，$P = 2.04 \times 10^{-8}$）也在先前的 GWAS 中被报道为精神分裂症的易感位点，并在最近研究中被确认了其介导精神分裂症的潜在通路[19]。本研究观察到的最强关联信号 DNAJC12 基因（*rs*7918428，$P = 1.82 \times 10^{-9}$），据报道也与早发性帕金森病、进行性神经发育迟缓和肌张力障碍有关[20]。最后，我们还发现了三个新的 AD 相关 SNP 位点，它们分别位于 7 号染色体（*rs*764009，$P = 2.23 \times 10^{-8}$）、2 号染色体（*rs*7593443，$P = 2.83 \times 10^{-8}$）和 1 号染色体（*rs*2170560，$P = 3.11 \times 10^{-8}$）。

此外，对于模型 M1 和 M2 发现的最显著的基因 LNX1（Chr4，*rs*2082768，M1：$P = 1.80 \times 10^{-6}$，M2：$P = 8.12 \times 10^{-7}$），也有相关的研究证实该基因在哺乳动物神经干细胞的分化、神经系统突触形成、神经传递和调节神经胶质功能等领域的重要作用[21]。Li 等[22]在一项近期的研究中发现 LNX1 缺陷会导致小鼠体内相应受体的功能减退，并观察到 LNX1 缺陷小鼠成年阶段的认知障碍。因此，虽然未达到统计学显著性，但是 M1 和 M2 提示的 LNX1 基因确实可能与人类的认知功能相关，具体的生物学机制有待进一步研究。

本文所提出方法的优点是可以在 GWAS 中处理带有缺失基因型的家族关联数据，同时可以用于校正人群结构，而无需家庭成员的基因型信息。该方法使用个体随机效应和家庭特异随机效应来控制遗传和非遗传因素带来的混杂，并能实现与使用 PCA 相似甚至更好的效果。通过计算机数值模拟研究发现，该方法在估计家族缺失基因型的联合概率分布以及估计 logistic 回归系数方面具有良好的表现。在实例分析中，我们的方法在发现遗传变异与表型之间的关联方面也表现出较好的优越性，不仅确认了已知的 AD 风险位点（例如 FRMD4A），而且新发现了达到全基因组显著水平的遗传变异风险位点，这些风险位点在仅使用先证者数据进行分析的两种方法中均未被发现。虽然一些位点先前没有被报道与 AD 相关，但是有研究显示它们与其他神经发育障碍有关。这些基因背后可能隐含着其他可能增加 AD 发病风险的生物学路径，为潜在的干预措施提供方向。在模型的扩展和推广方面，一方面我们可以利用该模型同时研究多个表型变量，如对阿尔茨海默病和帕金森病联合建模；另一方面，我们也可以进一步应用所提出的混合效应 logistic 回归模型在报告家庭成员表型信息的电子健康记录中绘制疾病的遗传变异图。此外，我们的模型或可允许根据先证者的生物标志物信息对其亲属进行个性化风险预测，协助进行 AD 的早期预防和疾病诊疗。

本研究仍存在一定的局限性：①我们仅在西班牙裔人群中进行了建模，研究样本量有限；②同样由于数据样本量限制，我们并没有对结果进行验证；③AD 的发病与多种遗传因素和非遗传因素密切相关[1, 23]，虽然本研究纳入了较多的基因信息，但是建模时仅考虑了年龄、性别和受教育程度三个人口学信息，这在一定程度上可能会影响检出结果的准确性。

综上所述，我们提出了一种混合效应 logistic 回归模型，有效的解决了家族关联数据 GWAS 分析亲属基因型未知的问题，为处理家族关联数据提供了新的方法参考。本文的研究发现可以为后续阿尔茨海默病遗传通路研究、疾病检测和治疗提供新的启发和思路。

参 考 文 献

[1] Jose A, Hector MGL, Gabriel CLG. Chapter 13-Alzheimer's disease [J]. Handbook of Clinical Neurology, 2019, 167: 231-255.

[2] Global, regional, and national burden of Alzheimer's disease and other dementias, 1990—2016: a systematic analysis for the Global Burden of Disease Study 2016[J]. Lancet Neurol, 2019, 18(1): 88-106.

[3] 任汝静, 殷鹏, 王志会, 等. 中国阿尔茨海默病报告 2021[J]. 诊断学理论与实践, 2021, 20(4): 21.

[4] Armstrong RA. Risk factors for Alzheimer's disease[J]. Folia Neuropathol, 2019, 57(2): 87-105.

[5] Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease[J]. Nat Genet, 2013, 45(12): 1452-1458.

[6] Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk[J]. Nat Genet, 2019, 51(3): 404-413.

[7] Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing[J]. Nat Genet, 2019, 51(3): 414-430.

[8] Marioni RE, Harris SE, Zhang Q, et al. GWAS on family history of Alzheimer's disease[J]. Transl Psychiatry, 2018, 8(1): 99.

（下转第 714 页）