doi:10. 3969/j.issn.1672-5565.2016.02.10

基于基因组关联数据识别阿尔茨海默病相关通路

周小禹
(广西省桂林市全州二中生物组,桂林市 541000)

摘　要:阿尔茨海默病又称老年性痴呆,是一种复杂的中枢神经系统退行性疾病,本文选取一套阿尔茨海默病全基因组关联分析(GWAS)数据,利用 ProxyGeneLD 软件进行基因水平上的检验,利用 WebGestalt 数据库进行遗传通路分析,识别出 320 个显著(P<0.05)的阿尔茨海默病相关基因、8 个显著的 KEGG 通路和 41 个显著的 GO 功能类,这些研究结果对进一步揭示阿尔茨海默病潜在的发病机制具有重要意义。

关键词:全基因组关联研究;遗传通路;阿尔茨海默病
中图分类号:R541　　**文献标志码**:A　　**文章编号**:1672-5565(2016)02-123-04

Identifying risk pathways of Alzheimer's disease based on the data of genome-wide association studies

ZHOU Xiaoyu
(*Quanzhou No.*2 *Middle School*, *Guilin* 540000, *China*)

Abstract: Alzheimer's disease (AD), which is also called senile dementia, is a kind of complex central nervous system degenerative diseases. In this paper, we selected a genome-wide association study dataset of AD, and conducted a gene-based test using ProxyGeneLD and a pathway analysis using WebGestalt. We identified 320 significant AD genes (P<0.05), 8 significant KEGG pathways and 41 significant GO pathways (P<0.05). These results are helpful to elucidate the potential pathogenies of Alzheimer's disease.

Keywords: Genome-wide association studies; Genetic pathways; Alzheimer's disease

阿尔茨海默病 (Alzheimer's disease, AD),又称老年性痴呆,是一种复杂的中枢神经系统退行性疾病,以高级认知功能障碍为特征,以老年斑、神经纤维缠结和神经元丢失为主要病理改变的综合病。阿尔茨海默病发病率与年龄呈现正相关性。据估计,65 岁老年人发病率为 4.4%,90 岁以上老年人的发病率是 22%[1]。随着世界人口日趋老龄化,阿尔茨海默病已经成为当前老年医学面临的最严峻的问题之一。在我国,人口的老年化进程不断加快,如何对这两种常见老年疾病进行有效地预防和早期治疗,已成为我国面临的一项关系到国家人口研究数据识别阿尔茨海默相关的风险位点和遗传通路。

目前,欧洲和美国研究人员普遍采用全基因组关联研究(Genome-Wide Association Studies, GWAS)和候选基因研究的方法筛选阿尔茨海默病易感基因,取得了前所未有的成就。一些新的阿尔茨海默病易感基因,例如 CR1, BIN1, CLU, PICALM, MS4A4/MS4A6E, CD2AP, CD33, EPHA1 和 ABCA7 等已经被逐渐报道[2]。AD 作为一种人类复杂性状,涉及多个基因,但是每个基因对表型只有较小或微小的影响,只有若干个基因共同作用,才可对表型产生明显影响,即个体表型是与多个基因相关的变异协同作用的结果。因此,利用生物信息学识别 AD 相关的遗传学通路,更能揭示潜在的遗传机制。本研究通过利用生物遗传通路分析方法,分析基因组关联研究 AD 数据,识别影响 AD 的生物学遗传通路,揭示 AD 潜在的遗传机制。

1　材料与方法

1.1　遗传关联研究数据

选取一套 AD GWAS 数据,该数据公开获得。

收稿日期:2016-03-31;**修回日期**:2016-05-25.

* **通信作者**:周小禹,男,中学教师,研究方向:生物技术及数据分析;E-mail:171393516@qq.com.

该数据包含11 789欧洲人,其中有3 941 AD 病例个体和7 848个对照个体。本套数据包含529 205个人类常染色体单核苷酸多态性(Single nucleotide polymorphism, SNP)数据。所有 SNP 信息来自人类基因组版本 37(Human genome build 37)和 SNP 版本 132(dbSNP build 132)。加性模型下的回归分析用来衡量单个 SNP 和 AD 的关联。最终,我们得到了 761 个 $P<=1.00\times10^{-3}$ 的 SNPs。更多详细信息在文章中有描述[3]。

1.2 方法

1.2.1 基因水平上检验 AD GWAS

利用 ProxyGeneLD 软件进行基因水平上的检验。该软件考虑到人类基因上复杂的连锁不平衡模式,利用 HapMap 数据的连锁不平衡信息[4],矫正由于基因长度所引起的显著性膨胀问题。如果有一些 SNP 在 HapMap 数据中高度连锁平衡 (r2>0.8),那么这些 SNP 归为一类,作为单个遗传信号。然后检验每个 GWAS 显著的 SNP 是否包含在某一个类里面。最后,每个基因都赋予一个经过多重检验矫正的 P 值。我们选取矫正后 P<0.05 的基因进行通路水平上的检验。

1.2.2 通路水平检验 AD GWAS

利用 WebGestalt 数据库进行遗传通路分析,连接地址为: http://bioi nfo. vanderbilt. edu/webgestalt/[5]。对于一个给定的通路,采用超几何分布检验识别这一通路与 AD 关联是否显著。在某一个通路中观测到 K 个 AD 相关基因的 P 值可以计算为:

$$P = 1 - \sum_{i=0}^{K} \frac{\binom{S}{i}\binom{N-S}{m-i}}{\binom{N}{m}}$$

N 是所有参考基因的数据, S 是所有 AD 相关基因的数目,m 是通路中含有的基因的数目,K 是通路中含有的 AD 相关基因的数目。我们采用 FDR(False discovery rate)方法进行多重检验校正[5]。对于任何一个通路,如果校正后的 P<0.05,并且含有至少 5 个 AD 相关基因,则认为这个通路和疾病显著相关。

2 结果与分析

2.1 显著富集的 KEGG 通路

利用 ProxyGeneLD 软件进行基因水平上的检验,识别出 320 个 AD 基因。根据 WebGestalt 进行通路水平上的检验,分别发现了 8 个显著的 KEGG 通路(P<0.05)。其中 Cell adhesion molecules,是最显著的遗传通路。其它通路主要包含 3 个心血管疾病通路(Dilated cardiomyopathy, Arrhythmogenic right ventricular cardiomyopathy 和 Hypertrophic cardiomyopathy),代谢通路(Glycosaminoglycan biosynthesis-chondroitin sulfate 和 Purine metabolism),神经系统和疾病(5 个),见表 1。

表 1 显著的 KEGG 通路

Table 1 The significant KEGG pathways

Pathway ID	Pathway Name	NG	NGR	rawP	adjP
Hsa04514	Cell adhesion molecules	7(309)	134(30 642)	4.36×10^{-4}	1.73×10^{-2}
Hsa05414	Dilated cardiomyopathy	6(309)	90(30 642)	3.07×10^{-4}	1.82×10^{-2}
Hsa05412	Arrhythmogenic right ventricular cardiomyopathy	5(309)	73(30 642)	8.66×10^{-4}	2.58×10^{-2}
Hsa04080	Neuroactive ligand-receptor interaction	9(309)	269(30 642)	1.74×10^{-3}	2.59×10^{-2}
Hsa05410	Hypertrophic cardiomyopathy	5(309)	83(30 642)	1.54×10^{-3}	2.62×10^{-2}
Hsa00532	Glycosaminoglycan biosynthesis - chondroitin sulfate	3(309)	22(30 642)	1.36×10^{-3}	2.69×10^{-2}
Hsa00230	Purine metabolism	7(309)	162(30 642)	1.34×10^{-3}	3.18×10^{-2}
Hsa05010	Alzheimer's disease	8(309)	164(30 642)	2.74×10^{-4}	3.26×10^{-2}

注:NG:输入基因列表中注释到通路中的基因数据 (输入基因数目); NGR:通路中还有的基因数目(参考基因中的所有基因数目); O:观测到在通路中与 AD 相关基因的数目,rawP:原始的 P 值,adjP:矫正后的 P 值。

Notes: NG means the number of inputting genes; NGR means all of the genes in reference gene set; O means the number of genes associated with AD in a pathway; rawP means the original P value; adjP means the P value after correction.

2.2 显著富集的 GO 通路

利用 320 个 AD 基因,我们得到了 41 个显著的功能类 P<0.05。我们进一步发现这些功能里都与代谢有关。主要包含 reverse cholesterol transport (GO:0043691), phospholipid efflux (GO:0033700), triglyceride homeostasis (GO:0070328), activation of phospholipase C activity (GO: 0007202), lipid metabolic process (GO:0006629), cholesterol efflux (GO: 0033344), cholesterol homeostasis (GO: 0042632), cholesterol metabolic process (GO:

0008203), lipid transport (GO: 0006869), 和 lipoprotein metabolic process (GO:0042157)。有趣的是, cell adhesion (GO: 0007155) 依然是第三显著的通路 $P=1.90\times10^{-5}$。本研究中我们列出前 10 个显著的功能类,见表 2。

表 2　前十个显著的 GO 通路

Table 2　The top 10 GO pathways

GO ID	GO Name	NG	NGR	rawP	adjP
GO:0043691	reverse cholesterol transport	6(309)	18(30 642)	1.68×10^{-8}	8.09×10^{-6}
GO:0007165	signal transduction	35(309)	1173(30 642)	1.19×10^{-8}	1.15×10^{-5}
GO:0007155	cell adhesion	22(309)	553(30 642)	5.93×10^{-8}	1.90×10^{-5}
GO:0007268	synaptic transmission	16(309)	372(30 642)	1.44×10^{-6}	3.46×10^{-4}
GO:0033700	phospholipid efflux	4(309)	10(30 642)	2.03×10^{-6}	3.91×10^{-4}
GO:0070328	triglyceride homeostasis	4(309)	16(30 642)	1.68×10^{-5}	2.69×10^{-3}
GO:0007271	synaptic transmission, cholinergic	4(309)	17(30 642)	2.18×10^{-5}	2.62×10^{-3}
GO:0007202	activation of phospholipase C activity	6(309)	58(30 642)	2.61×10^{-5}	2.79×10^{-3}
GO:0006629	lipid metabolic process	11(309)	247(30 642)	4.67×10^{-5}	3.46×10^{-3}
GO:0006810	transport	18(309)	607(30 642)	5.20×10^{-5}	3.57×10^{-3}

注:NG:输入基因列表中注释到通路中的基因数据 (输入基因数目); NGR:通路中还有的基因数目(参考基因中的所有基因数目); O:观测到在通路中与 AD 相关基因的数目,rawP:原始的 P 值,adjP:矫正后的 P 值。

Notes:NG means the number of inputting genes; NGR means all of the genes in reference gene set; O means the number of genes associated with AD in a pathway; rawP means the original P value; adjP means the P value after correction.

3　讨论与结论

生物信息学是生命科学、计算机科学和信息科学等学科逐步发展相互渗透的新兴交叉学科。随着对人类基因组计划的深入研究,生物信息学得到了蓬勃的发展,尤其是在了解各类疾病的发生机制及遗传基础上发挥了重要作用[6]。通过识别出与疾病发生发展相关的基因和通路,再据此进行实验验证,将是一种高效的研究途径。AD 是一种复杂疾病,利用生物信息学识别 AD 相关的遗传学通路,更能揭示 AD 潜在的遗传机制。本研究我们利用生物信息学方法,采用生物遗传通路分析了一套 AD 全基因组关联研究数据。

本研究中,我们利用 ProxyGeneLD 软件进行基因水平上的检验,检测出 320 个显著($P<0.05$)的 AD 基因。利用 WebGestalt 进行通路水平上的检验,发现了 8 个显著的 KEGG 通路和 41 个显著的 GO 功能类($P<0.05$)。其中,我们发现 Cell adhesion molecules 是 KEGG 中最显著的遗传通路,也是 GO 通路中第三显著的遗传信号。我们查阅了相关文献,前期的研究支持了我们的发现。Lambert 和 Jones 等人都使用 ALIGATOR 和 GenGen 软件,并且都用来分析两套 AD GWAS 数据,但是这两个研究却没有产生一致的结果[7]。中国科学院的研究人员认为不同的研究可能有共享的遗传通路。为了检验这种假设,他们应用多重遗传通路分析方法,分析了来自法国和美国的 AD GWAS 数据(9 580个样本)。在 KEGG 数据库中,发现了一个与阿尔茨海默病高度相关的遗传通路(Cell adhesion molecules, CAM)。在 GO 数据库,他们重复了这一发现[7]。进一步我们发现,cell adhesion molecules 还参与了好多自身免疫疾病[8]。

同时,我们发现了 AD 参与了 3 条直接与心血管病相关通路,Dilated cardiomyopathy, Arrhythmogenic right ventricular cardiomyopathy 和 Hypertrophic cardiomyopathy。该结果也进一步验证了早期广东医科大学的发现。研究人员对来自欧洲的14 138个样本(6 399 个 AD 疾病个体和 7 739 个对照个体) 进行了全基因组范围内基于基因和生物遗传通路水平上的分析。利用基因水平上关联检验的方法得到了 1 458 个显著($P<0.05$)的 AD 基因。然后采用生物遗传通路分析对 1 458 个 AD 相关基因进行 KEGG 和 GO 遗传通路注释,结果发现了 3 个与心血管疾病有关的显著富集的 KEGG 通路:viral myocarditis (hsa05416), dilated cardiomyopathy(DCM)(hsa05414),hypertrophic cardiomyopathy (HCM) (hsa05410)。因此,本研究中,我们验证支持了 dilated cardiomyopathy (DCM) (hsa05414) 和 hypertrophic cardiomyopathy (HCM) (hsa05410)。

本研究中,进一步发现了显著的代谢 GO 功能类,进一步支持了早期的研究结果。国外的研究人

员 Jones 等分析了两套 AD GWAS 数据. 识别出 25 个显著的 GO 功能类,大部分都与代谢有关[6]。本研究中,我们选用 KEGG 通路和 GO 通路,主要基于以下考虑:KEGG 是通过人工文献阅读和提取的生物学知识数据库,没有明显的分层迭代结构[9];GO 数据库主要是基于计算预测以及人工注释,具有明显的分层迭代结构,而且 GO 分析假定每个 GO 功能分类是彼此独立的,只有大约 1%的功能分类是经过试验验证的[10]。因此,这两个数据库形成了很好的补充。

尽管本研究得到了有价值的结果,但仍有其局限性。例如本研究中我们采用了多重检验校正,但是还不足以校正所有的偏倚,研究结果最好需要随机扰动试验。但是目前我们无法获得原始的基因型数据,因此我们后期的研究中还需要获得原始基因型数据,来进一步验证研究结果。

参考文献

[1] BETTENS K, SLEEGERS K, BROECKHOVEN C V. Current status on Alzheimer disease molecular genetics: from past, to present, to future[J]. Human Molecular Genetics, 2010,19(R1):R4-R11.

[2] BERTRAM L, MCQUEEN M B, MULLIN K, et al. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database[J]. Nature Genetics, 2007,39(1):17-23.

[3] SEGRÈ A V, CONSORTIUM D, INVESTIGATORS M, et al. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits[J]. Plos Genetics, 2010,6(8):e1001058.

[4] CAPONE R, JANG H, KOTLER S A, et al. Probing structural features of Alzheimer's amyloid-beta pores in bilayers using site-specific amino acid substitutions[J]. Biochemistry, 2012, 51(3): 776-785.

[5] ZHANG B, KIROV S, SNODDY J. WebGestalt: an integrated system for exploring gene sets in various biological contexts[J]. Nucleic Acids Research, 2005, 33(Web Server issue): W741-748.

[6] YOONA H, FLORES L F, KIM J. MicroRNAs in brain cholesterol metabolism and their implications for Alzheimer's disease[J]. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 2016, doi:10.1016/j.bbalip.2016.04.020.

[7] LIU G, JIANG Y, WANG P, et al. Cell adhesion molecules contribute to Alzheimer's disease: multiple pathway analyses of two genome-wide association studies[J]. Journal of Neurochemistry, 2012, 120(1):190-198.

[8] LIU G, JIANG Y, CHEN X, et al. Measles contributes to rheumatoid arthritis: evidence from pathway and network analyses of genome-wide association studies[J]. PLoS One, 2013,8(10):e75951.

[9] JOZWIAK K, ZEKANOWSKI C, FILIPEK S. Linear patterns of Alzheimer's disease mutations along alpha-helices of presenilins as a tool for PS-1 model construction[J]. Journal of Neurochemistry, 2006,98(5):1560-1572.

[10] SIVAPRAKASAM K. Towards a unifying hypothesis of Alzheimer's disease: cholinergic system linked to plaques, tangles and neuroinflammation [J]. Current Medicinal Chemistry, 2006,13(18):2179-2188.