·学习·交流·发现·

全基因组关联研究理论及其在流行病中的应用

卢昕 阚飙 刘剑锋 张勤

【关键词】 全基因组关联研究；流行病

Theory of genome-wide association study and its application in epidemic diseases

LU Xin[1,2], *KAN Biao*[1], *LIU Jian-feng*[2], *ZHANG Qin*[2]. 1 *State Key Laboratory for Infectious Disease Prevention and Control, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing* 102206, China; 2 *Key Laboratory Animal Genetics and Breeding of the Ministry of Agricultural, State Key Laboratory of AgroBiotechnology, College of Animal Science and Technology, China Agricultural University*

Corresponding author: ZHANG Qin, Email: qzhang@cau.edu.cn

This work was supported by grants from the National Basic Research Program (973 Program) (No. 2006CB102104), *National Natural Science Foundation* (No. 30972092), *National High Technology Research and Development Program (863 Program)* (No. 2008AA101002) *and Important National Science and Technology Specific Projects* (No. 2008ZX10004-009, 2009ZX10601) *of China.*

【Key words】 Genome-wide association study; Epidemic diseases

全基因组关联研究(GWAS)利用遍布于全基因组范围的分子标记(主要为单核苷酸多态性，SNP)并借助强大统计学工具，对影响某些复杂性状如疾病易感的遗传变异进行鉴定和分析。

一、基本原理

1. 统计学分析原理：同经典的病例对照研究相似，GWAS假设SNP与疾病发生关联，则理论上病例中该SNP的等位基因频率应不同于对照组[1]，然后通过检验进行验证。目前，主要分为两种实验设计：基于无关个体或基于家系(Family-based association)。

2. 分子标记：GWAS的一大特点就是不再需要选择候选基因或染色体区域，而是针对全基因组中所有的SNP进行分析。在人类疾病研究中GWAS的大量应用得益于人类基因组图谱的破译和单倍型图谱(HapMap)的构建。目前，最新的HapMap数据库(http://hapmap.ncbi.nlm.nih.gov/)于2010年8月公布Ⅰ、Ⅱ、Ⅲ期数据的合并结果，在数据库中犹他州的欧洲西部和北部的后裔(CEU)、中国北京汉族人和日本东京人(CHB+JPT)以及尼日利亚约巴鲁人(YRI)3个种族人群各确定大约4 M个SNP，在其余7个亚种或亚群中也各发现约1.4 M个SNP。最新的千人基因组(http://www.1000genomes.org)研究结果中，共确认约15 M个SNP，预计包含人类SNP(MAF≥1%)总数的95%左右，超过dbSNP数据库(约11 M个SNP)。目前成熟的基因芯片产品中SNP密度一般均在1 M个以上，已基本可以满足研究的需要。但是，所有的芯片均不能覆盖所有的变异，在研究中可以利用那些与直接分型的SNP处于连锁不平衡的标记来提高对变异的检测能力。主要有两种方法：一种是基于单倍型分析的方法，另一种是填充法(Imputation methods)，通过填充法，一般可以将芯片上的SNP增加至2倍左右，如Affymetrix 6.0分型芯片，经质控后约可得到0.8 M个SNP，结合HapMap Ⅱ信息，可填充至约2.4 M个，如果再利用最新的千人基因组SNP数据库，其填充的效率将更为惊人。

3. 假阳性校正：多重假设检验导致的Ⅰ型错误增加和假阳性关联是GWAS面临的重要问题之一。多重假设检验的次数取决于所选全基因组SNP的数量。有多种方法可以校正关联研究中多重假设检验后的P值以减少假阳性结果，如Bonferroni校正法、递减调整法、数据重排法、控制错误发现率法等[2]。除此之外，假阳性率还会受到来自系统偏差的影响，其中最主要的是群体分层效应。即使研究群体源自于同一种族，该问题也仍然可能存在，这主要是由于亚人群混杂造成的[3]。要减少亚人群混杂的影响，常采用的手段是加大样本量，并尽可能选择遗传均质的群体(如出生地、年龄结构、种族、性别比例相同)；也有人采用核心家系设计和传递不平衡检验(TDT)分析方法，可以有效避免群体分层的影响[4]；针对无关个体构成的群体，常采用基因组对照(genome control, GC)[5]、结构关联(structured association, SA)[6]和主成

DOI: 10.3760/cma.j.issn.0254-6450.2011.10.020

基金项目：国家重点基础研究发展计划(973计划)(2006CB102104)；国家自然科学基金(30972092)；国家高技术研究发展计划(863计划)(2008AA101002)；国家重大科技专项(2008ZX10004-009, 2009ZX10601)

作者单位：102206 北京，中国疾病预防控制中心传染病预防控制所传染病预防控制国家重点实验室(卢昕、阚飙)；中国农业大学动物科技学院农业生物技术国家重点实验室农业部畜禽遗传育种重点开放实验室(卢昕、刘剑锋、张勤)

通信作者：张勤，Email: qzhang@cau.edu.cn

分分析(principal component analysis, PCA)[7]方法消除群体分层。

4. 重复验证:单纯依靠调整P值无法判断SNP与性状是否真实关联,有必要进行验证[8],目前多采用多阶段研究设计进行验证。

5. 公共数据信息平台建设:全球范围内越来越多的研究机构建立公用的GWAS数据信息平台。美国国立卫生院癌症研究所的癌症易感性遗传标记计划(Cancer Genetic Markers of Susceptibility Initiative, CGEMS)率先将乳腺癌和前列腺癌GWAS结果(包括P值、RR值及其95%CI)公开。此外,糖尿病遗传学计划(Diabetes Genetics Initiative, DGI)的研究者们也公开了研究数据。美国国立卫生院更是制定了相关的政策督促更多的研究组共享数据。对于GWAS分析结果,目前,最权威的数据库为美国NIH的GWAS数据库(http://www.genome.gov/26525384)。该网站收录的GWAS结果均经过严格筛选,其中收录的研究均需对100 000个以上的SNP位点进行分析,且不包含候选基因法,研究所得到的显著SNP与性状的相关度P值需$<1.0\times10^{-5}$。截止到2011年4月13日,该数据库里共收录862篇人类GWAS文章报道的4305个SNP位点。

6. 局限性:GWAS所采用的统计学方法尚不足够成熟,假阳性结果不可避免,所以需要更大样本量更多人群去进行重复验证,而且GWAS所确定的遗传变异的具体致病机制也有待于更多实验去验证。目前通过GWAS所发现的复杂疾病致病基因仅可以阐明与流行病有最强关联的基因变异,仍有更多的基因和它们之间的相互作用机制有待深入挖掘,甚至还要理清这些基因与环境之间的相互作用机制。而且,GWAS难以检测罕见变异,所以难以检测到那些最小等位基因频率(MAF)在5%以下的变异所产生的效应。目前很多GWAS发现的许多显著SNP位点位于非基因编码区,对于这一现象还没有很合理的解释,但这些区域很可能与基因表达调控或蛋白质修饰有关。无论如何,GWAS开创了一个流行病致病机制研究的新阶段。

二、GWAS在流行病中的应用

GWAS的兴起使得研究人员能够从全基因组范围去寻找相关的序列变异,筛选出与疾病性状关联的SNP,从而在对疾病遗传基础及致病机制的阐明中迈进一大步。2005年,Klein等[9]在*Science*发表一篇有关年龄性黄斑变性症(AMD)的GWAS文章,这是首次较为系统地应用GWAS方法对复杂疾病开展研究。该研究对96名患者和50名对照进行全基因组扫描,采用大约11万个SNP标记,发现与该疾病相关的1个SNP位点rs380309(P=0.0043),并在随后的研究中发现补体因子F基因(CFH)与该位点存在着连锁不平衡。对CFH基因进行重测序后在其外显子中发现与该疾病紧密关联的1个突变。该研究的成功为后来的GWAS提供很好的借鉴。

2009年,Le Clerc等[10]利用病例对照分析法,对85名病例和2049名对照进行研究,重点对HIV载量进行分析,发现转化因子β通路在AIDS中有着很重要的作用。Rauch等[11]采用GWAS方法比较1015名慢性肝炎患者和347名自愈者的SNP位点,结果显示白细胞介素28B(interleukin 28B, IL28B)对机体控制肝炎病毒起着重要作用。McGovern等[12]针对896名节段性肠炎患者和3204名健康者进行GWAS,发现岩藻糖转移酶基因(FUT2)与节段性肠炎显著关联。2009年,Kamatani等[13]用两阶段法对乙型肝炎进行GWAS研究,发现HLA-DP中存在1个与乙型肝炎易感性显著相关的位点。Duerr等[14]对溃疡性结肠炎进行GWAS,发现IL23R与溃疡性结肠炎显著相关,该基因将可能成为该病治疗靶位点。2009年,Zhang等[15]采用两阶段GWAS策略去研究麻风病的易感性问题,第一阶段采集706名患者和1225名健康人,发现93个与麻风病易感性显著相关的位点;第二阶段采集3254名患者和5955名健康人进行验证,最终发现核苷酸结合寡聚蛋白2(NOD2)介导的信号通路在麻风病易感中起着重要的作用。Jallow等[16]采用GWAS对2500名儿童疟疾进行分析,并用3400名儿童进行第二阶段的重复验证。有3个研究小组分别采用GWAS策略去研究对于丙型肝炎病毒(HCV)的治疗反应,发现IL28B在病毒的阻碍中起着很重要的作用[17-19]。2008—2010年就有6篇文章对溃疡性结肠炎进行GWAS,发现一些可能的易感基因[20],另有10篇文章采用GWAS研究风湿性关节炎,发现一些风湿性关节炎易感基因(IL6ST、SPRED2、RBPJ、CCR6、IRF5和PXK等)。2008年,Weidinger等[21]对1530名个体的353 569个SNP进行GWAS,发现IgE Fc片段受体1A基因(Fc fragment of IgE, FCER1A)为慢性肝病易感基因之一(表1)。

三、讨论

随着GWAS的逐步深入,人们逐渐认识到,越复杂的性状(如人类大多数流行病),往往越需要大量的样本来重复验证。而要想实现大样本研究的相互重复,数据共享是必经之路。大量的共享数据,对发现更多微效的与复杂疾病关联的基因变异、阐明基因变异与环境因素之间的交互作用关系至关重要。复杂性状应当是基因与环境复杂地相互作用的产物,数据的不断积累与共享,将为更复杂的模型如基于通路的模型(pathway-based model)、上位模型(epistatic model)等提供足够的检测效力,从而更深刻精细地揭示其机制。除了在统计学上找到更多关联的结果,并确保其准确可靠外,GWAS还需要深入全面揭示基因或遗传标记的功能,以及环境因素对遗传因素的影响。复杂性状的产生或疾病的发生,应当是一个异常复杂的过程。各种水平的遗传因素如DNA、RNA、蛋白质、代谢物等,同各种水平的外部环境因素如居住地、职业、教育水平、生活习惯等,以及各种水平的内部环境因素如性别、年龄、身高、体重等,相互作用,交织成网。有必要充分运用各种生物信息学的方法来对人体这个复杂网络进行研究。

参考文献

[1] Hardy J, Singleton A. Genomewide association studies and human disease. N Engl J Med, 2009, 360(17): 1759-1768.

表1 利用GWAS对人类流行病进行研究实例分析

研究性状	SNP技术平台(SNP数)	起始实验	重复实验	文献出处
AIDS病毒载量	Illumina(291 119)	85/2049	未进行重复实验	[10]
乙型肝炎	Illumina(499 544)	179/934	1599/2821[a]; 308/546[b]	[13]
慢性丙型肝炎感染	Illumina(~500 000)	1015/347	未进行重复实验	[11]
节段性肠炎	Illumina(304 825)	896/3204	1174/357	[12]
溃疡性结肠炎	Illumina(308 332)	547/548	401/433[c]	[14]
麻风病	Illumina(491 883)	706/1225	3254/5955	[15]
疟疾	Affymetrix(402 814)	958/1382	1087/2376	[16]
对于丙型肝炎的治疗反应	Illumina(311 159)	131/162	261/294	[17]
	Affymetrix(621 220)	72/82	122/50	[18]
	Illumina(565 759)	571/566[d]	未进行重复实验	[19]
溃疡性结肠炎	Illumina(513 923)	376/934	376/1097	[20]

注:[a]日本人群;[b]泰国人群;[c]除此之外,还包括1个核心家庭设计;[d]包含高加索人、非洲裔美国人、西班牙人;表中分子为病例人数,分母为对照人数

[2] Yan WL. Genome-wide association study on complex diseases: genetic statistical issues. Hereditas, 2008, 30(5): 543-549. (in Chinese)
严卫丽. 复杂疾病全基因组关联研究进展——遗传统计分析. 遗传, 2008, 30(5): 543-549.

[3] Newton-Cheh C, Hirschhorn J. Genetic association studies of complex traits: design and analysis issues. Mutat Res, 2005, 573(1-2): 54-69.

[4] Bacanu S, Devlin B, Roeder K. The power of genomic control. Am J Hum Genet, 2000, 66(6): 1933-1944.

[5] Devlin B, Roeder K. Genomic control for association studies. Biometrics, 1999, 55(4): 997-1004.

[6] Pritchard J, Rosenberg N. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet, 1999, 65(1): 220-228.

[7] Price A, Patterson N, Plenge R. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet, 2006, 38(8): 904-909.

[8] Chanock SJ, Manolio T, Boehnke M. Replicating genotype-phenotype associations. Nature, 2007, 447(7145): 655-660.

[9] Klein RJ, Zeiss C, Chew EY. Complement factor H polymorphism in age-related macular degeneration. Science, 2005, 308(5720): 385-389.

[10] Le Clerc S, Limou S, Coulonges C, et al. Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). J Infect Dis, 2009, 200(8): 1194-1201.

[11] Rauch A, Kutalik Z, Descombes P, et al. Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study. Gastroenterology, 2010, 138(4): 1338-1345.

[12] McGovern DP, Jones MR, Taylor KD, et al. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Hum Mol Genet, 2010, 19(17): 3468-3476.

[13] Kamatani Y, Wattanapokayakit S, Ochi H, et al. A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. Nat Genet, 2009, 41(5): 591-595.

[14] Duerr RH, Taylor KD, Brant SR, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science, 2006, 314(5804): 1461-1463.

[15] Zhang FR, Huang W, Chen SM, et al. Genomewide association study of leprosy. N Engl J Med, 2009, 361(27): 2609-2618.

[16] Jallow M, Teo YY, Small KS, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet, 2009, 41(6): 657-665.

[17] Suppiah V, Moldovan M, Ahlenstiel G, et al. IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. Nat Genet, 2009, 41(10): 1100-1104.

[18] Tanaka Y, Nishida N, Sugiyama M, et al. Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. Nat Genet, 2009, 41(10): 1105-1109.

[19] Ge D, Fellay J, Thompson AJ, et al. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature, 2009, 461(7262): 399-401.

[20] Asano K, Matsushita T, Umeno J, et al. A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. Nat Genet, 2009, 41(12): 1325-1329.

[21] Weidinger S, Gieger C, Rodriguez E, et al. Genome-wide scan on total serum IgE levels identifies FCER1A as novel susceptibility locus. PLoS Genet, 2008, 4(8): e1000166.

(收稿日期:2011-03-22)
(本文编辑:万玉立)