遗传 HEREDITAS (Beijing) 2007 年 1 月, 29(1): 3－7
ISSN 0253-9772 www.chinagene.cn

综 述

DOI: 10.1360/yc-007-0003

人类复杂疾病关联研究中群体分层的检出和校正

智联腾[1,2], 周钢桥[1,2], 贺福初[1,2]

1. 军事医学科学院放射与辐射医学研究所, 基因组学与蛋白质组学研究室, 北京 100850;
2. 北京蛋白质组研究中心, 功能基因组学研究室, 北京 102206

摘要: 病例对照研究是鉴定多基因疾病易感位点重要的遗传流行病学方法, 而群体分层是导致病例对照研究关联研究结果出现偏倚甚至是假关联的重要原因之一。文章对人群分层的检出及校正的方法和原理进行了阐述, 包括基于核心家系的传递/不平衡检验(TDT)以及基于不相关基因组遗传标记的基因组对照(GC)和结构化关联(SA)等, 并且对这几种方法进行了比较。

关键词: 关联研究; 人群分层; 传递不平衡检验; 基因组对照; 结构化关联

Detection and controlling for population stratification in association studies of human complex disease

ZHI Lian-Teng[1,2], ZHOU Gang-Qiao[1,2], HE Fu-Chu[1,2]

1. *Department of Genomics & Proteomics*, *Beijing Institute of Radiation Medicine*, *Beijing* 100850, *China*;
2. *Department of Functional Genomics*, *Beijing Proteome Research Center*, *Beijing* 102206, *China*

Abstract: Case-control studies, which serve as standard design for genetic association analysis, can be the most practical and powerful approach to detect genetic polymorphisms contributing to susceptibility to complex human diseases. However, considerable concern has been expressed that this approach is prone to population stratification, which can lead to biased or spurious results. We review several methods to detect and account for population stratification; these methods include nuclear family-based transmission/disequilibrium test (TDT), and genomic control (GC) and structured association (SA) based on unlinked genetic markers.

Keywords: association study; population stratification; transmission/disequilibrium test; genomic control; structured association

近年来, 多种人类复杂疾病(Polygenic disease, Complex disease), 如哮喘、心血管疾病、原发性高血压、精神分裂症和糖尿病等的发病率呈现出明显的上升趋势。从现代的遗传学观点看, 人类复杂疾病的发生和发展是由特定的环境因素、机体本身的遗传因素(主要是遗传易感性, Genetic susceptibility)以及上述两类因素间协同交互作用的结果。近年来, 随着国际人类基因组计划、单核苷酸多态性(Single nucleotide polymorphism, SNP)计划和单倍型图谱(Haplotype map, HapMap)计划等的相继实施, 机体

收稿日期: 2006-03-17; 修回日期: 2006-09-07
基金项目: 中国高技术研究发展计划(编号: 2001AA224011)、科技专项计划(编号: 2002BA711A10)、军队医药卫生科研基金(编号: 01Z018)和国家自然科学基金委创新研究群体科学基金(编号: 30321003)资助[Supported in part by grants from the Chinese High-tech Program (No.2001AA224011 and 2002BA711A10), Medicine and Health Research Program (No.01Z018) and Chinese National Science Fund for Creative Research Groups (No.0321003)]
作者简介: 智联腾(1977－), 男, 山西省太谷县人, 在读博士研究生, 专业方向: 医学遗传学与基因组学。Tel: 010-80727777-1226; E-mail: zhilt@126.com
通讯作者: 贺福初(1962－), 男, 湖南人, 博士, 研究员, 中国科学院院士, 研究方向: 基因组学与蛋白质组学。E-mail: hefc@nic.bmi.ac.cn

本身的遗传因素对疾病发生和发展的重要作用越来越受到人们的重视，并已成为后基因组时代基因组学研究的热点领域之一[1]。鉴定人类复杂疾病的遗传易感位点或基因、阐明其遗传学基础，不仅是对疾病病因学的发展和完善，也是实现群体和个体疾病风险预测、个体化医疗及新药开发的关键环节[1~3]。

人类复杂疾病遗传易感基因定位研究主要有连锁分析(Linkage analysis)和关联分析(Association study)两种方法。众所周知，连锁分析在孟德尔遗传的、具有明显主基因效应的单基因遗传病(Monogenic disease)的致病基因的定位方面取得了极大的成功。然而，由于受遗传异质性、基因-基因和基因-环境相互作用、外显不全和拟表型等因素的影响，使得连锁分析难以在复杂疾病易感基因的定位研究中发挥理想的作用。在此情形下，关联研究方法重新得到人们的重视。关联分析是在群体水平上研究某种疾病与某个特定等位基因的频率相关性，最常见的实验设计方法是病例对照研究(Case control study)。它以某人群内一组患有某种疾病的人群(称为病例组)和同一人群内未患该病但在与患病有关的某些已知因素(包括社会人口学因素和环境暴露)方面与病例组相似的人群(称为对照组)作为研究对象，通过比较病例和对照组间遗传标记频率的差异，从而推断该标记与该疾病易感性的相关关系；如果遗传标记的等位频率在病例组和对照组间具有显著的统计学差异，则可认为该等位型与疾病存在统计学关联，并可推断该标记存在于疾病易感基因座内，或者与疾病易感基因间存在连锁不平衡关系。它无需家系资料，避免了家系患病成员临床和人口学资料和DNA标本不易获取等限制因素。因此关联分析更灵活、使用范围更广泛，是检测等位基因与疾病之间非随机关联以及进行基因作图的有效工具，在复杂疾病易感基因的定位研究和药物基因组学研究等诸多领域皆有广泛的应用[4]。

关联研究固然有其优越性，但也存在一定的局限性，其中最主要的局限是关联结果易受混杂因素(Confounding factor)的干扰[5~7]。在诸多混杂因素中，一个极其重要但又常常被忽视的因素即是群体的遗传分层现象(Population stratification)。在病例对照研究的实验设计中，如果某遗传标记的等位频率在病例和对照组间存在显著差异，但这个遗传标记并不与疾病表型相关，则认为该研究中存在群体分层现象。群体分层往往是由于遗传背景不一的亚人群混合所致，其产生机制复杂，可能与各亚人群祖先的迁移模式、婚配习惯、生殖强弱及基因组的随机突变等因素有关。群体分层对遗传关联分析的直接影响是可能导致结果偏倚，产生假阳性或假阴性的结果。正因如此，很多关联研究的结果难以在不同人群的研究中得到重复。如，在美国黑人和高加索人中高血压的发病率有很大差异，在黑人中流行比例要更高，因此在利用黑人和白人的混合人群进行关联研究时，那些在黑人中频率较高的等位基因型将有可能被认为与疾病相关联，虽然这些等位型可能实际上与该疾病毫不相关[8]。因此在进行病例对照的关联研究时，尽可能排除群体分层的干扰显得尤为重要。

排除群体分层的干扰，常采用的手段是加大样本量并尽可能的选择一个遗传上相对均质的群体，如出生地、年龄结构、种族、性别比例相同，人口流动性小的隔离群体。然而，即便如此，仍不能彻底排除可能存在的隐藏分层的干扰。因而，有的研究者倾向于采用基于家系研究的策略，即选择以受累家系为基础的内在对照组(由双亲或同胞组成)，采用传递/不平衡检验(Transmission/disequilibrium test, TDT)等方法进行检验[9,10]。该方法有效地解决了上述难题，但同时也引出了一些新的问题，如统计效能(Statistics power)较低等。为此，近年来又发展了一些新的方法，如选取一定数目的、与疾病无关的遗传标记在病例对照样本中同时进行分型(Genotyping)，并以这些数据为基础，应用专门的软件或算法推断群体分层存在与否以及分层的程度，并对关联分析的结果进行校正[11,12]。

1 基于核心家系的研究策略

该方法采用患者的生物学亲属作为虚拟“对照”。由于患者和生物学亲属间具有相似的遗传背景，因而可以有效的避免群体分层的影响[13,14]。比较常用的分析方法是TDT分析：从有一个受累子女的核心家系(Nuclear family)中抽样，以双亲未传递给子女的基因型为虚拟“对照”，通过研究杂合子双亲传递遗传标记等位基因给受累子女的传递率，比较传递与不传递之间的差异，以确定存在关联的情况下是否连锁。这种方法评价杂合子双亲在传递变异等位基因给发病子女时，其传递率是否偏离服从孟德尔遗传规律且没有连锁时的期望传递率。

TDT 分析不需要疾病传递的遗传模型(复杂疾

病往往很难确定遗传模式); 既可以用于仅含单个发病子女的小家系, 也可以用于有多个发病成员的大家系; 不受群体分层的影响, 因此有效的避免了由此产生的虚假关联。而且采用该方法只需对候选标记进行分析, 不需要对其他不相关的标记位点进行分析。其缺点是: (1)不管是在对定性性状还是定量性状的研究中, 其统计效能均低于病例对照研究设计; (2)当没有关联存在时, 不能用来检测疾病与遗传标记是否连锁; (3)只有遗传标记为杂合子的双亲的基因型信息才可用于 TDT 分析(对于一个 SNP 位点, 杂合度的最大值为 50%, 因此每次分析中, 至少一半的双亲资料被排除); (4)该分析中一对病例对照需要分析 3 个样品(先证者及其双亲), 因此分型效率仅是所有入选样本的三分之二, 不如关联研究经济; (4)对于晚发疾病(患者双亲的 DNA 可能难以获得)在研究的实施上存在很大的难度。由此可见, TDT 分析虽然解决了人群分层的干扰, 但也存在一些缺陷。因此, 是否采用基于核心家系的 TDT 研究策略, 将很大程度上取决于所研究的疾病的特征, 若核心家系比较容易收集, 提倡采用该方法, 反之, 基于群体的病例对照研究仍是关联研究的最佳选择。当然, 对群体病例对照研究形成某一假说后继之以 TDT 在核心家系或同胞对中进行验证可能是未来关联研究的发展趋势。

2　基于基因组对照的研究策略

如上所述, 基于群体的关联研究设计不仅在统计效能上强于基于核心家系的研究, 而且更为经济、可操作性更强。近年来发展了一些可以有效控制病例对照研究中群体分层对关联结果的影响的新方法[5,11,12,15], 这些方法均能在存在群体分层的情况下进行病例对照的关联研究, 同时, 对于原本设计为无群体分层的病例对照样本, 该方法又可定量检测事实上存在的群体分层的程度[15]。

由于受到遗传和进化等多方面因素的影响, 基因组中任意一个遗传标记的等位频率在遗传背景不同的种族中可能不尽相同。因此, 在进行基因型-表现型的相关性研究中, 如果对照组和病例组来自遗传背景不相同的人群, 完全可以根据从基因组中随机选取的中性遗传标记等位频率的差异而检测出群体分层, 并继而对群体分层进行校正。

目前, 应用随机选取的中性遗传标记进行群体分层的校正主要有两种途径, 一是 Devlin 等人提出的"基因组对照"(Genomic control, GC)方法[16,17], 二是由 Pritchard 和 Rosenberg 提出的"结构化关联"(Structured association, SA)方法[18~21]。

2.1　基因组对照(GC)

GC法采用卡方检验(或趋势检验)进行统计学分析, 该方法不考虑环境因素对疾病的直接作用, 并假设群体分层的效应在基因组水平上是一个常量, 即选择若干不相关的遗传标记分别进行病例对照卡方检验时, 若人群一定, 则卡方值应是一个常量。在病例对照研究中, 群体分层将导致χ^2检验统计量发生偏倚, 并且引起错误的拒绝零假设(即导致检验统计量的零分布发生波动或膨胀)。采用GC的方法, 可以有效的检出这类偏差[14]。Devlin等人采用不相关的遗传标记估算出的偏离系数λ(或称膨胀因子, Inflation factor)来估计出可能的分层效应, 并且进一步用估算出的λ去除χ^2值(即χ^2/λ), 以达到校正的目的。偏离系数代表了人群分层的程度, 与人群分层呈正相关[17]。

根据所选取的不相关的遗传标记的数据, 可用两种方法估算偏离系数: (1) 采用参数途径进行估计[17,22], 该方法是一个理想化的、纯统计的算法。具体做法是取各个卡方检验χ^2的中值, 按如下公式计算: $\lambda=\text{Median}\ (Y_i^{\,2})/0.456$。式中$Y_i^{\,2}$指第$i$个遗传标记在病例对照研究中的卡方值; 0.456 是自由度为 1 时, 卡方值理论上的中值。(2) 采用模拟的方法, 结合95%可信区间(Confidence interval, *CI*)的上限进行估计。该方法估计的偏离系数为各个遗传标记χ^2值的平均值乘以相应的系数, 选取的位点数目不同系数也不同, 而且一般都大于 1, 因此该方法估计的偏离系数较保守。此外, 对于严重分层的病例对照群体, 该方法无能力检出[15]。随机选取的不相关的遗传标记数一般为几十个; 同时, 若候选标记等位频率小于 15%, 在随机选取不相关标记时还需进行频率匹配, 以减少偏差。

采用GC法检测群体分层的统计效能将会随着样本量的增大而提高, 因此一些中等程度的分层情况在大的样本量中更易检测得到。该方法比较方便、灵活, 但校正后的结果相对过于保守。对于近期混合人群, GC法不适用[23]。另外, 要求这些作为"指示剂"的遗传标记是双等位型的, 而且必须是随机选取、彼此互不相关, 与被研究的位点之间不应存在连锁不平衡关系, 符合Hardy-Weinberg平衡定律,

一般要求杂合度大于 40%，标记所在基因组区域不受自然选择(Positive selection)的作用，等等。与TDT分析相比较，当群体分层效应实际并不存在时，GC的统计效能要强于TDT；当分层事实上存在的情况下，TDT的统计效能强于GC [14]。

Hao等[22]选取了 10,000 个SNPs位点，分别在四个人群(包括 1 个亚洲人群、一个非洲裔美国人群和 2 个高加索人群)中进行了分型。基于这些数据，以两种不同的模式——不同的遗传标记数和不同的人群分层程度——评价GC法的性能。结果表明，当分层事实上存在时，两种模式都可以用 20~50 个随机的SNP位点检出人群分层，混合的个体将被分散在同类的亚人群；分层程度不是很明显时，两种方法都不敏感，但由分层引起的偏倚还是可以得到校正。由此可见，通过中性遗传标记推测的膨胀因子能够有效的校正由于人群分层导致的混杂效应。

2.2 结构化关联(SA)

SA也是采用随机选取的遗传标记来检测分层，与前者不同的是，它不是通过利用这些遗传标记检测检验统计量的分布、估计偏离系数，而是通过估计亚人群的数目以及各个个体应属于哪个亚人群来进行校正[23]。假设研究所选的病例对照群体为遗传异质性较高的人群，由几个同源的亚人群组成，则利用基因组内多个遗传标记的分布特点，将研究群体中在遗传上同类的个体重新分配到亚人群中，使各个亚人群内病例组和对照组间的匹配更佳，并在各个亚人群中分别进行疾病-标记的关联研究。最后，总体的关联研究结果是各个亚人群各自关联研究的综合。该方法利用全部所选遗传标记的数据来检测分层的存在，因此最大程度的利用了各个遗传标记的分型数据[3,23]。

SA采用基于模型的贝叶斯聚类算法，假设每个亚人群都是一个参数模型，每个亚人群可通过各个遗传标记的一系列频率来区分，标记可以是微卫星多态、RFLP，也可以是SNP，这是与GC法的区别之一。通过对多个遗传标记的分型数据的分析，对群体结构进行推断，假设共有K个亚人群(K未知)，这样每个个体可通过概率估计，归类到各个亚人群中。假如基因型显示某些个体是混合型，也可能归到两个或多个亚人群中，因此对于近期混合的人群同样适用，这也是SA法与GC法的区别之一。该模型并不需要假定特定的突变进程，常见的大部分突变只要互相不关联、不是紧密连锁且在人群内遵循Hardy- Weinberg平衡的均可以用于该模型中。而且只要选择合适数量的位点，该方法都有较高的准确性[23~25]。另外，该方法有很强的遗传学基础，并且考虑到了自然选择的作用，这也是与GC法的不同之处。

采用SA方法进行亚人群的推定，其估算的准确性依赖于采用的算法、人群分层的程度以及数据量等。目前对于K的推断主要有两种算法：一种是基于贝叶斯框架(Bayesian framework)的算法，并采用马尔可夫蒙特卡罗(Markov chain Monte Carlo, MCMC)法从亚人群等位频率及个体祖先等的联合后验分布中抽样。该算法可将各种先验信息如某些个体的种族资料等直接整合，使得在遗传上相似的个体被确定为属于同一个亚人群[18,19]。采用该算法校正群体分层可以通过在线运行*structure*和*STRAT*程序来完成(www.stats.ox.ac.uk/mathgen/software. html)。另一种是基于最大期望(Expectation Maximum, EM)算法的最大似然法[23,24]。K的估计采用基于AIC (Akaike information criterion)准则的最大罚分似然估计，其中罚分随着模型中参数的增加而增加。该算法与其他方法不同的是，它的出发点不是进行显著性检验，而是估计候选位点效应的大小，而且其前提是假设该效应可以在所有的亚人群中采用相同的参数进行模拟。另外该方法不允许亚人群间存在混杂。总之，在亚人群信息很充分时，采用上述两种方法都可以有效地对亚人群进行估算；但是，在大样本极限时，第二种方法容易导致对亚人群的估计过头。

SA法与GC法相比较，当候选位点在不同人群中效应相同时，GC统计效能略高于SA；然而，由于SA法允许不同亚人群中等位型的效应不同，因此在这种情况下SA法的统计效能强于GC法。另外在可用的遗传标记较少或推断人群分层有一定困难时，GC法有时会更好些。与TDT法相比较，当样本总数相同时(即R个病例与R个对照的关联研究和 2R/3 个核心家系的研究)，假设等位效应在所有亚人群中都相同时，GC法与SA法的统计效能均强于TDT，且GC法强于SA法；若与R个核心家系的分析相比较，则TDT的效能稍强[23,26]。

总之，多数研究者都已认识到群体分层对于关联研究结果的影响，而且已经提出一些对群体分层进行校正的方法，研究者可选择适合自己的校正方法，也可以几种方法联用来进行校正。但由于各种方法均有各自的优缺点，而且采用这些方法的研究

报道并不多, 因此这些新方法的统计学效能及有效性还需通过实际应用及通过与其他方法进行比较才能最终确定。

参考文献(References):

[1] Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. Mapping complex disease loci in whole-genome association studies. *Nature*, 2004, 429: 446－452.[DOI]

[2] Suh Y, Vijg J. SNP discovery in associating genetic variation with human disease phenotypes. *Mutat Res*, 2005, 573(1-2): 41－53.

[3] Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet*, 2003, 361: 598~604.[DOP]

[4] Keavney B. Genetic association studies in complex diseases. *J Hum Hypertens*, 2000, 14: 361－367.[DOI]

[5] Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM. Control of confounding of genetic associations in stratified populations. *Am J Hum Genet*, 2003, 72: 1492－1504.[DOI]

[6] Khlat M, Cazes MH, Genin E, Guiguet M. Robustness of case-control studies of genetic factors to population stratification: magnitude of bias and type I error. *Cancer Epidemiology, Biomarkers Prev*, 2004, 13: 1660－1664.

[7] Marchini J, Cardon LR, Phillips M S, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet*, 2004, 36: 512－517.[DOI]

[8] Lander ES, Schork NJ. Genetic dissection of complex traits. *Science*, 1994, 265: 2037－2048.[DOI]

[9] Zhao HY. Family-based association studies. *Stat Methods Med Res*, 2000, 9: 563－587.[DOI]

[10] Hinds DA, Stokowski RP, Patil N, Konvicka K, Kershenobich D, Cox DR, Ballinger DG. Matching strategies for genetic association studies in structured populations. *Am J Hum Genet*, 2004, 74: 317－325.[DOI]

[11] Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet,* 2003, 33: 228－237.[DOI]

[12] Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet,* 1999, 65: 220－228.[DOI]

[13] Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet,* 1993, 52: 506－513.

[14] Bacanu SA, Devlin B, Roeder K. The power of genomic control. *Am J Hum Genet*, 2000, 66: 1933－1944.[DOI]

[15] Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol*, 2001, 20: 4－16.[DOI]

[16] Devlin B, Roeder K. Genomic control for association studies. *Biometrics*, 1999, 55: 997－1004.[DOI]

[17] Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Pop Biol*, 2001, 60: 155－166.[DOI]

[18] Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured population. *Am J Hum Genet*, 2000, 67: 170－181.[DOI]

[19] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*, 2000, 155: 945－959.

[20] Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 2003, 164: 1567－1587.

[21] Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky L A, Feldman M W. Genetic structure of human populations. *Science*, 2002, 298: 2381－2385.[DOI]

[22] Hao K, Li C, Rosenow C, Wong W H. Detect and adjust for population stratification in population-based association study using genomic control markers: an application of Affymetrix Genechip Human Mapping 10 K array. *Eur J Hum Genet*, 2004, 12: 1001－1006.[DOI]

[23] Pritchard JK, Donnelly P. Case-control studies of association in structured or admixtured populations. *Theor Pop Biol*, 2001, 60, 227－237.[DOI]

[24] Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet*, 2001, 68: 466－477.[DOI]

[25] Tang H, Quertermous T, Rodriguez B, Kardia SL, Zhu X, Brown A, Pankow JS, Province MA, Hunt SC, Boerwinkle E, Schork NJ, Risch NJ. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am J Hum Genet*, 2005, 76: 268－275.[DOI]

[26] Zhu X, Zhang SL, Zhao HY, Cooper RS. Association mapping using a mixture model for complex traits. *Genet Epidemiol*, 2002, 23: 181－196.[DOI]