·基础理论与方法·

全基因组关联研究中的二阶段病例对照设计

马昭君　易洪刚　赵杨　陈峰

【导读】 全基因组关联研究(GWAS)已成为寻找疾病致病基因的重要手段，但是研究费用昂贵，大部分研究者选择了资源利用率更高的二阶段设计。为系统阐述二阶段病例对照设计的研究设计和统计分析方法，论文作者结合实例分析介绍了其设计的基本原理，在GWAS中的实施步骤、统计分析策略以及应用特点和现状。

【关键词】 全基因组关联研究；二阶段设计；病例对照

Using two-stage case-control designs to study the genome-wide association *MA Zhao-jun, YI Hong- gang, ZHAO Yang, CHEN Feng. Department of Epidemiology and Biostatistics, School of Public Health, Nanjing Medical University, Nanjing* 210029, China
Corresponding author: CHEN Feng, Email: fengchen@njmu.edu.cn
This work was supported by a grant from the National Natural Science Foundation of China (No. 81072389) *and Natiural Science Foundation of Higher Education Institutions of Jiangsu Province* (No. 10KJA330034).

【Introduction】 Genome-wide association study is an important approach to identify common genetic variants that predispose to human disease. Because of the high cost of genotyping on hundreds of thousands of markers on thousands of subjects, a more cost-effective two-stage case-control design is applied by most genome-wide association studies. To describe the design and statistical methods of the two-stage case-control study, this paper introduces the principles of two-stage case-control design, its implementing steps in genome-wide association study and the features of its application. The method is illustrated with an example.

【Key words】 Genome-wide association study; Two-stage design; Case-control

随着各种高通量技术(大规模基因组测序技术、基因芯片和质谱技术)的快速发展，全基因组关联研究(genome-wide association study, GWAS)的应用领域越来越广阔，越来越多地被应用于复杂性疾病如肿瘤、糖尿病等研究。GWAS中关心的三个主要问题分别是检验效能、假阳性和经费。要控制假阳性，用最少的经费达到预定的检验效能，需要良好的研究设计。GWAS的研究设计目前主要分为单阶段设计和二阶段设计。单阶段病例对照设计是选择了足够的病例和对照样本后，对所有研究对象全基因组中的单核苷酸多态性(SNP)进行基因分型。该设计需要较大的样本含量。一般GWAS所研究的SNP达到几十万甚至上百万个，因此这种研究耗资巨大，且由于许多本来在实验早期就可以确定的与疾病无关的DNA片段也被包括进来，从而使得该设计资源利用率很差[1]，造成一定的浪费。

Sobell等1993年首次提出在寻找致病基因的关联研究中使用二阶段病例对照设计[2]。2002年，Satagopan等[1]首次系统阐述了二阶段病例-对照设计的使用。随后，该设计方法不断得到完善，使其更适用于候选基因研究，甚至扩展应用到GWAS中[2-8]。2006年，波士顿大学医学院联合哈佛大学等报道的关于肥胖的研究中，首次将二阶段设计应用于GWAS[9]。为了节约研究经费，提高资源利用率，越来越多的研究者在复杂性疾病致病基因的关联研究中采用了二阶段设计。

本研究主要介绍二阶段病例对照设计在GWAS中的应用。

基本原理

1. 实施步骤：第一阶段选择n_1个病例对照样本，在该样本中对全基因组范围内选择的所有m_1个SNP进行基因分型。然后分析每个SNP与疾病的关联，分别计算*OR*值及其检验统计量。按照检验统计量的绝对值从大到小排列全部SNP，选择其中位于前

DOI: 10.3760/cma.j.issn.0254-6450.2010.10.025
基金项目：国家自然科学基金(81072389)；江苏省高校自然科学研究重大项目(10KJA330034)
作者单位：210029 南京医科大学流行病与卫生统计学系
通信作者：陈峰，Email: fengchen@njmu.edu.cn

$\pi_{markers}$%的SNP，也就是说，选择前面的$m_2=m_1\pi_{markers}$%个SNP。第二阶段选择另一部分独立的病例对照样本n_2个，在该样本中只对第一阶段所选出的m_2个SNP进行基因分型。然后对第二阶段的结果进行统计分析或者结合两个阶段的结果进行统计分析，寻找与疾病相关联的阳性SNP位点。

2. 统计学分析：在二阶段设计提出的初期，第二阶段的数据被看作是第一阶段的重复，故第二阶段的数据被单独进行统计分析（χ^2检验或logistic回归），将有统计学意义的位点作为与疾病相关的位点。2006年Skol等[10]发现，将二个阶段的数据联合后进行统计分析的检验效能比单独将第二阶段作为重复阶段分析的检验效能更高，此后更多的研究者选择对二阶段设计的GWAS资料进行联合分析。

二阶段设计的联合分析方法：假设GWAS全部研究样本含量为n，病例与对照样本含量相等，所要研究的全部SNP为m，$\pi_{samples}$为第一阶段所用样本含量占全部样本含量的比例，则第一阶段样本含量$n_1=n\times\pi_{samples}$，第二阶段样本含量$n_2=n\times(1-\pi_{samples})$。$\hat{p}_1'$和$\hat{p}_1$分别表示病例和对照中的致病等位基因频率。则检验统计量可记作

$$z_1=\frac{\hat{p}_1'-\hat{p}_1}{\sqrt{[\hat{p}_1'(1-\hat{p}_1')+\hat{p}_1(1-\hat{p}_1)]/(n\cdot\pi_{samples})}} \quad (1)$$

在无关联的零假设下，第一阶段样本含量较大时，z_1服从均数为0，方差为1的标准正态分布。可以用标准正态分布的性质来确定一个界值C_1，作为筛选进入第二阶段SNP的标准，这样，$P(|z_1|>C_1)=\pi_{markers}$，$\pi_{markers}$表示全部$m$个SNP中进入第二阶段的比例。如果将第二阶段作为重复阶段分析，检验统计量z_2的公式和z_1相似。在联合分析中，要产生一个新的统计量，该统计量在两阶段之间存在异质性时也适用：

$$z_{joint}=\sqrt{\pi_{samples}}\,z_1+\sqrt{(1-\pi_{samples})}\,z_2 \quad (2)$$

将求出的检验统计量z_{joint}与界值C_{joint}比较，如果$z_{joint}>C_{joint}$，说明等位基因频率在病例和对照中不同，该位点与所研究疾病有相关性。Skol提供了一个在各种m，n，$\pi_{markers}$，$\pi_{samples}$条件下计算C_1，C_2，C_{joint}的软件（http://csg.sph.umich.edu）。

3. 单阶段设计与二阶段设计的比较：

（1）检验效能：假设GWAS无经费限制，选择单阶段病例对照设计可达到最大检验效能。但是，一个GWAS通常要研究几十万甚至上百万个SNP，要将所有个体的全部SNP基因分型，费用非常昂贵。

Skol等[6]和Muller等[8]的理论研究及模拟实验结果表明，在GWAS中应用二阶段病例对照设计，在大量减少研究经费的同时，可以保证研究的检验效能。

第一、二阶段每个SNP基因分型的费用分别记为t_1，t_2，$r=t_2/t_1$。在不同的等位基因频率AF（allele frequency）、基因型相对风险GRR（genotype relative risk）、r的情况下，将最佳的二阶段设计与单阶段设计进行比较（表1）。本研究中最佳的二阶段设计指的是在控制Ⅰ类错误、检验效能不变的前提下，使研究的经费达到最低的设计。

表1 单阶段设计与二阶段设计比较

r	AF	GRR	n	$\pi_{samples}$(%)	T_{two}/T_{one}(%)[a]
100	0.1	1.25	15 200	59	63.10
		1.5	4 000	63	66.70
		2	1 220	60	64.00
	0.3	1.25	6 500	59	63.10
		1.5	1 900	60	64.00
		2	620	62	65.80
	0.5	1.25	5 740	59	63.10
		1.5	1 740	63	66.70
		2	610	62	65.80
200	0.1	1.25	15 200	59	67.20
		1.5	4 000	63	70.40
		2	1 220	60	68.00
	0.3	1.25	6 500	59	67.20
		1.5	1 900	60	68.00
		2	620	62	69.60
	0.5	1.25	5 740	59	67.20
		1.5	1 740	63	70.40
		2	610	62	69.60

注：[a]二阶段设计与单阶段设计的研究经费相比的百分比；表中结果的前提条件包括，$n_{case}=n_{control}$，检验效能$1-\beta=0.9$，$m=500\,000$，$\pi_{markers}=0.1\%$，单个SNP假设检验经Bonferroni法调整后的$\alpha=1\times10^{-7}$

第一阶段基因分型一般采用商业芯片，按照国内目前GWAS相关费用，1 M芯片价格约2300元，每个SNP平均0.0023元；第二阶段选择部分SNP进行基因分型，需定制特定芯片，若选择1000～5000个SNP，每个SNP费用约0.5元，则$r=t_2/t_1=0.5/0.0023\approx217$，若选择的SNP 5000以上，每个SNP基因分型费用会有所下降，因此本文中r设置为100和200。当$r=100$或200时，在不同的AF、GRR组合下，二阶段设计都可以达到与单阶段设计相同的检验效能，同时可以节约30%～40%的研究经费。

（2）控制假阳性错误：二阶段病例对照设计中的第二阶段，是将第一阶段分析结果中与所研究疾病有关联的SNP或者是可能有关联的SNP，在另外一个独立样本中进行基因分型，以验证这些SNP与所研究疾病是否仍有关联。因此，第二阶段也可以看

为第一阶段结果的验证阶段，可以起到控制假阳性错误的作用。

在二阶段设计的GWAS中，第一阶段作为探索阶段，主要是筛选SNP，样本含量往往较少，因此第一阶段的检验效能较低。研究者为了避免漏掉与疾病相关的SNP，在选择进入第二阶段的SNP时，会将检验水准α_1设置得较大，此时第一阶段总的Ⅰ型错误率会增大。如某GWAS总的SNP数m_1=500 000，若设置第一阶段$\alpha_1=10^{-5}$，则总的Ⅰ型错误率$=1-(1-10^{-5})^{500\,000}=0.993$。因此，第二阶段作为验证阶段，需要设置较严格的检验水准α_2，以控制整个研究的假阳性率。如进入第二阶段的SNP个数$m_2=m_1\times\pi_{markers}$，则$\alpha_2=0.05/(m_1\times\pi_{markers})$。同时，由于第二阶段增大了样本含量，保证了研究的检验效能。

PubMed中检索到2009年关于肿瘤的GWAS文献共发表24篇，其中19篇应用病例对照研究。19篇GWAS中有18篇应用的是二阶段或三阶段病例对照设计。有3篇未提供第一阶段假设检验的具体结果，在其余的15篇文献中，8篇通过第二阶段的重复验证排除了第一阶段的部分阳性结果[11,12]。由此可见，二阶段病例对照设计与单阶段设计相比，可以更好地控制研究的假阳性。

(3)研究经费：假设一个GWAS全部研究样本含量为n，且病例与对照样本含量相等，所要研究的全部SNP为m_1，其中真正的致病基因位点为$D(\geqslant 1)$个，则其余m_1-D个位点与疾病无关。关联研究的目的是从这些真正的致病基因位点中至少识别出d个($1\leqslant d\leqslant D$)。

单阶段设计：将n个个体的m_1个SNP都进行基因分型，每一个SNP都作假设检验，例如logistic回归，然后挑选出检验统计量最大的d个SNP，推断其为致病基因位点。

设每个SNP基因分型的费用为t_1，则该设计方法基因分型所需经费为

$$T_1=n\times m_1\times t_1 \tag{3}$$

二阶段设计：第一阶段对n_1个个体的全部m_1个SNP进行基因分型，并计算其与疾病关联的检验统计量，定义$\pi_{samples}$(<1)为第一阶段所用样本含量占全部样本含量的比例，则

$$n_1=n\times\pi_{samples} \tag{4}$$

按照检验统计量排序后，根据其顺位挑选最大的一部分SNP，定义该部分的比例为$\pi_{markers}$($0<\pi_{markers}<1$)。第二阶段，在另外的$n_2=n-n_1$个个体中，对第一阶段所挑选的$m_2=m_1\times\pi_{markers}$个SNP进行基因分型。设第二阶段每个SNP基因分型的费用为t_2，则该设计方法基因分型所需经费为

$$T=n\,\pi_{sample}m_1t_1+n(1-\pi_{sample})m_1\pi_{markers}t_2 \tag{5}$$

在$n=8000$，$m_1=1\,000\,000$，$\pi_{markers}=0.1\%$，按照现阶段$m_1t_1=$￥2300，$t_2=$￥0.5的价格，根据公式3计算，单阶段设计研究经费$T_1=n\times m_1\times t_1=$￥1840(万元)。在上述条件不变时，由公式5可见，二阶段设计将大大降低研究经费。如$\pi_{sample}=50\%$时，$T=$￥1120(万元)，而且随着π_{sample}的减小，T呈明显的下降趋势。

实例分析

2009年*Journal of Human Genetics*报道了一篇关于鼻咽癌的GWAS文章，研究者采用二阶段病例对照设计，寻找与鼻咽癌有关联的致病基因位点[13]。

第一阶段：研究者对马来西亚华人中的111例鼻咽癌患者和260例对照的554 496个SNP进行基因分型。去除21 448个基因分型缺失率>0.02、或MAF<0.05的SNP，对其余SNP采用Fisher确切概率法进行关联分析，结果显示533 048个SNP均$P>1\times10^{-5}$。

第二阶段：根据第一阶段的关联分析结果，选择P值最小的200个SNP进入研究的第二阶段，在另外一组独立的病例对照样本中对其进行基因分型，包括268例病例和252例对照。对两阶段数据进行联合分析，结果显示3p21染色体的ITGA9基因中的位点rs2212020与鼻咽癌易感性有相关性($P=8.27\times10^{-7}$，$OR=2.24$)。选择rs2212020附近40 kb连锁不平衡区域的19个SNP进行基因分型。联合分析结果显示共有8个SNP的$P<1\times10^{-5}$，其中rs189897的P值最小，确认其与鼻咽癌易感性之间的关联有统计学意义($P=6.85\times10^{-8}$，$OR=3.18$)。

该研究如果以第一阶段全基因组扫描每个样品$m_1t_1=$\$500，第二阶段每个基因手工分型每个样品每个SNP $t_2=$\$0.07的价格，按照公式5计算，基因分型的费用大约为19.35万美元，若应用单阶段设计，则基因分型费用将达到44.55万美元，二阶段设计与单阶段设计比较，可节约基因分型费用56.57%。

讨　论

随着芯片技术的发展，全基因组扫描的价格不断下降，但是对大量样本的几十万或上百万SNP进行分析的GWAS，其高额的研究经费仍然不是一般实验室所能承受的。因此大部分的GWAS的研究者

选择采用资源利用率更高的二阶段病例对照设计。

目前大部分研究者的研究结果表明，第一阶段样本含量占全部样本含量的比例$\pi_{samples}$在50%左右，进入第二阶段研究的SNP占全部SNP比例$\pi_{markers}$在1%左右时，二阶段设计的检验效能(power)与单阶段设计较为接近。

但是第一阶段和第二阶段的样本含量如何分配能达到最佳的二阶段设计，还取决于在第二阶段研究每个SNP的费用与在第一阶段研究每个SNP的费用之比[5]。这里"最佳的二阶段设计"指的是在GWAS中保证研究的检验效能的同时使研究经费达到最低。

事实上，大多的研究在开始之前已经确定了所要研究的样本含量n。当n固定，所要研究的SNP个数m给定时，检验效能最高的仍然是将全部n个个体的全部m个SNP基因分型的单阶段设计，即使最佳的二阶段设计仍会损失一小部分的检验效能。当然，检验效能的损失可以<1%，而节约三分之一以上的研究经费。如果条件允许，检验效能的部分损失还可以通过增加部分样本含量来弥补[8]。

在二阶段设计中，如何选择一部分SNP进入GWAS的第二阶段有很多种方法。常见的是根据第一阶段基因分型结果，分析每个SNP与疾病的关联，按照检验统计量的绝对值排列全部SNP，选择其中位于前$\pi_{markers}$%的SNP。该方法简单易行，但是可能会漏选一部分可能与疾病有关联的有研究价值的SNP。

Rao[14]提出一种复杂但较全面的选择方法，首先根据单个SNP与疾病关联分析的P值来选择进入第二阶段研究的大部分SNP，然后利用已有(自己或他人研究中已经发现某些区域与疾病连锁)的连锁信息来选择剩下的小部分SNP，尽管这些SNP在第一阶段的关联分析中P值没有达到进入第二阶段的标准。

由于二阶段设计对资源利用率更高，因此在GWAS中得到广泛使用，在某些大样本量的GWAS中，研究者开始采用三阶段病例对照设计，其基本原理和二阶段设计相似，是将第二阶段的过程延续为两个阶段，其进入第二阶段研究的SNP更多，在第二阶段研究结束后根据前两阶段研究结果选择一少部分SNP进入第三阶段，在独立的样本中进行分析。甚至某些研究者在GWAS中采用了四阶段病例对照设计[15]，这种多阶段的病例对照设计方法需要更进一步的研究和讨论。

参 考 文 献

[1] Satagopan JM, Verbel DA, Venkatraman ES, et al. Two-stage designs for gene-disease association studies. Biometrics, 2002, 58:163-170.

[2] Thomas DC, Xie RR, Gebregziabher M. Two-stage sampling designs for gene association studies. Genet Epidemiol, 2004, 27: 401-414.

[3] Satagopan JM, Venkatraman ES, Begg CB. Two-stage designs for gene-disease association studies with sample size constraints. Biometrics, 2004, 60:589-597.

[4] Zuo YJ, Zou GH, Zhao HY. Two-stage designs in case-control association analysis. Genetics, 2006, 173:1747-1760.

[5] Wang HS, Thomas DC, Pe'er I, et al. Optimal two-stage genotyping designs for genome-wide association scans. Genet Epidemiol, 2006, 30:356-368.

[6] Skol AD, Scott LJ, Abecasis GR, et al. Optimal designs for two-stage genome-wide association studies. Genet Epidemiol, 2007, 31:776-788.

[7] Lin DY. Evaluating statistical significance in twostage genomewide association studies. Am J Human Genet, 2006, 78:505-509.

[8] Muller HH, Pahl R, Schafer H. Including sampling and phenotyping costs into the optimization of two stage designs for genomewide association studies. Genet Epidemiol, 2007, 31:844-852.

[9] Herbert A, Gerry NP, McQueen MB, et al. A common genetic variant is associated with adult and childhood obesity. Science, 2006, 312(5771):279-283.

[10] Skol AD, Scott LJ, Abecasis GR, et al. Joint analysis is more efficient than replicationbased analysis for two-stage genome-wide association studies. Nature Genet, 2006, 38:209-213.

[11] Gudmundsson J, Sulem P, Gudbjartsson DF, et al. Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. Nature Genet, 2009, 41:460-464.

[12] Shete S, Hosking FJ, Robertson LB, et al. Genome-wide association study identifies five susceptibility loci for glioma. Nature Genet, 2009, 41:899-904.

[13] Ng CC, Yew PY, Puah SM, et al. A genome-wide association study identifies *ITCA9* conferring risk of nasopharyngeal carcinoma. J Human Genet, 2009, 54:392-397.

[14] Rao DC. An overview of the genetic dissection of complex traits. Adv Genet, 2008, 60:3-34.

[15] Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nature Genet, 2009, 41:334-341.

(收稿日期:2010-03-17)

(本文编辑:张林东)