·基础理论与方法·

基于基因水平主成分logistic回归模型在全基因组关联研究中的应用

易洪刚　沃红梅　赵杨　张汝阳　柏建岭　魏永越　陈峰

【导读】　探讨基于基因水平的主成分logistic回归模型分析方法及其在全基因组关联研究中的应用。以全基因组关联研究基因型模拟数据为例，介绍基于主成分的logistic回归模型在基因水平检测遗传变异与复杂性疾病之间关联的分析策略。模拟结果表明致病位点所在基因假设检验的P值在所有基因检验结果中为最小。研究结果提示在全基因组关联研究中，采用基于基因水平的主成分logistic回归模型一方面能够降低检验的自由度，另一方面能够处理单核苷酸多态性之间相关性问题，在检测致病基因与疾病关联时具有一定的效能。

【关键词】　主成分分析；Logistic回归模型；全基因组关联研究；关联

Gene-based principal component logistic regression model and its application on genome-wide association study　*YI Hong-gang, WO Hong-mei, ZHAO Yang, ZHANG Ru-yang, BAI Jian-ling, WEI Yong-yue, 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 grants from the National Natural Science Foundation of China (No. 81072389, 30901232), *Natural Science Foundation of Higher Education Institutions of Jiangsu Province* (No. 10KJA330034), *Specialized Research Fund for the Doctoral Program of Higher Education of China* (No. 20113234110002) *and the Priority Academic Program Development of Jiangsu Higher Education Institutions.*

【Introduction】 To explore the gene-based principal component logistic regression model and its application in genome-wide association study. Using the simulated genome-wide single nucleotide polymorphism (SNPs) genotypes data, we proposed a practical statistical analysis strategy—'the principal component logistic regression model', based on the gene levels to assess the association between genetic variations and complex diseases. The simulation results showed that the P value of genes in related diseases was the smallest among the results from all the genes. The results of simulation indicated that not only it could reduce the degree of freedom through hypothesis testing but could also better understand the correlations between SNPs. The gene-based principal component logistic regression model seemed to have certain statistical power for testing the association between genetic genes and diseases in the genome-wide association studies.

【Key words】 Principal component analysis; Logistic regression; Genome-wide association study; Association

全基因组关联研究(GWAS)已成为当前人类复杂性疾病致病机制研究最主要的研究手段之一。随着生物技术的不断发展，高通量检测技术已被广泛采用，目前GWAS所应用的高通量基因芯片至少能检测100万个单核苷酸多态性(SNPs)。但GWAS的样本量却相对小得多(如2000～10 000例)。因此，GWAS就会存在高维数据的"维度灾难"(curse of dimensionality)问题，即变量(SNPs或者基因)数非常多，而同时样本例数相对非常少[1]。此时，许多传统的统计方法如logistic回归模型直接用于GWAS高维数据分析将面临诸多问题，如多重共线性(multicollinearity)、多重比较(multiple comparisons)等问题。

主成分分析(PCA)作为一种降维方法，已被广泛应用于高维数据分析中[2]。该方法主要寻找原始

DOI: 10.3760/cma.j.issn.0254-6450.2012.06.018

基金项目：国家自然科学基金(81072389)；国家自然科学基金青年基金(30901232)；江苏省高校自然科学研究重大项目(10KJA330034)；高等学校博士学科点专项科研基金(20113234110002)；江苏高校优势学科建设工程项目

作者单位：210029 南京医科大学公共卫生学院流行病与卫生统计学系

通信作者：陈峰，Email: fengchen@njmu.edu.cn

数据的线性组合，即用主成分有效表达原始数据的信息。各主成分间互相独立，第一主成分(PC_1)具有最大的方差，解释了原始数据最大的变异信息。基于主成分的回归模型相对于基于原始数据的分析由于降低了维度，从而增加高维数据分析的检验效能。近年来，主成分回归模型开始应用于GWAS的研究，相对于基于单个位点的分析方法，取得了较好的效果。本研究主要介绍基于基因水平主成分logistic回归模型在GWAS中的建模策略和应用。

基本原理

假设采用病例对照设计，样本中含病例(患病个体)和无关联的对照(未患病的个体)。对于每个个体，用$g_1, g_2, \cdots, g_k$表示某个区域(如基因)中K个SNPs基因型。采用相加模式，用0，1和2来表示第k个SNP($k=1, 2, \cdots, K$)中最小等位基因(minor alleles)的个数。用D表示疾病状态($D=0$表示对照，$D=1$表示患者)。为检测某个位点与疾病的关联，传统的logistic回归模型为：

$$\text{logit}[\Pr(D=1 \mid g_1, g_2, \cdots, g_K)] = \beta_0 + \beta_1 g_1 + \beta_2 g_2 + \cdots + \beta_K g_K \quad (1)$$

采用自由度为K的似然比检验(likelihood ratio test, LRT)来检验$H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$。

本研究采用基于基因水平的主成分logistic回归模型进行分析。首先，基于生物学先验信息，将SNPs划分成具有生物学功能的分析集合(SNPs set)。其次，在每个SNPs集合中进行PCA，通过主成分来提取、综合该区域内的基因型遗传信息。最后，在每个SNPs集合中选择一定数量的主成分，并建立主成分和疾病之间是否存在关联的logistic回归模型。具体方法和步骤：

(1)划分SNPs集合：根据生物学先验信息，划分全基因组中SNPs，形成有生物学意义的SNPs集合。理论上，SNPs集合可采用任意的划分方式，例如基因、单倍域(haplotype block)、基因通路等。根据基因进行划分是最常用的一种策略，将位于同一个基因中(或相邻区域)的SNPs划分成一个SNPs集合。例如，在Illumina HumanHap 500的芯片中，530 000个SNPs对应着大约17 800个基因，因此可将这些SNPs划分成大约17 800个SNPs集合，每个集合对应着某个基因(或其相邻区域)[3]。

(2)计算和选择主成分：基于每个SNPs集合基因型数据的相关系数矩阵，计算和选择主成分。在PCA中，主成分$PC_1, PC_2, \cdots, PC_k$为SNPs基因型数据的线性组合：

$$\begin{aligned} PC_1 &= e_1 g = e_{11} g_1 + e_{12} g_2 + \cdots + e_{1K} g_K \\ PC_2 &= e_2 g = e_{21} g_1 + e_{22} g_2 + \cdots + e_{2K} g_K \\ &\vdots \\ PC_K &= e_K g = e_{K1} g_1 + e_{K2} g_2 + \cdots + e_{KK} g_K \end{aligned} \quad (2)$$

式中，$e_i (i=1,2,\cdots,K)$是指在$e_i^T e_i = 1$条件限制下，使得每个主成分方差极大化的向量，并且不同主成分之间的方差协方差矩阵为0，即各个主成分之间互相独立。通过提取主成分的方法，将原来有相关性即连锁不平衡(linkage disequilibrium, LD)的K个SNPs转化成K个相互独立的主成分。其中，PC_1具有最大的方差，解释了该基因区域内最大的遗传变异信息，PC_2方差其次，以此类推。

设R为SNPs基因型数据的相关系数矩阵，则e_K为R的特征向量，$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K$为特征根，且$\text{Var}(PC_K) = \lambda_K$。则有

$$\sum_{k=1}^{K} \text{Var}(g_K) = \sum_{k=1}^{K} \text{Var}(PC_K) = \lambda_1 + \lambda_2 + \cdots + \lambda_K \quad (3)$$

第i个主成分能够解释的方差比例或贡献为$\lambda_i/(\lambda_1 + \lambda_2 + \cdots + \lambda_K)$。

(3)建立主成分logistic回归模型：将式(1)中的SNPs基因型数据替换成PC_K，建立基于基因水平的主成分logistic回归模型

$$\text{logit}[\Pr(D=1 \mid PC_1, PC_2, \cdots, PC_K)] = \beta_0 + \beta_1 PC_1 + \cdots + \beta_K PC_K \quad (4)$$

由于各个主成分均是按照其方差的大小进行排序的，因此，选择能够解释大部分变异信息的部分主成分($PC_1, PC_2, \cdots, PC_S, S<K$)来建模，采用自由度为$S$的LRT检验$H_0: \beta_1 = \beta_2 = \cdots = \beta_S = 0$。

实例分析

本研究采用病例对照设计，使用HapGen软件[4]产生人类基因组GWAS基因型数据，并基于基因水平采用主成分logistic回归模型进行分析。为了显示结果的稳定性，分析过程进行了100次的重复模拟，并以第22号染色体结果为例介绍分析步骤和结果。

本研究采用文献[5-7]的方法模拟产生1000个病例和1000个无关联对照的基因型数据。即基于国际人类基因组单体型图计划(the International HapMap Project, HapMap, http://snp.cshl.org/)数据(rel#22 - NCBI Build 36)中JPT+CHB人群模拟产生人类第22号染色体的SNPs基因型数据，染色体物理位置区域为14.43～49.58 Mb，包含32 668个SNPs。随机选择其中位于38.98 Mb的一个SNPs

(rs12484776)作为致病位点，该位点处于TNRC6B基因的内含子区域(intronic)，最小等位基因频率(minor allele frequency)设为15%。致病等位基因杂合子致病的优势比(odds ratio，*OR*)设为1.22，采用相乘遗传模式。疾病患病率设为15%。

首先，划分SNPs集合。基于PubMed中dbSNP数据库(Database of Single Nucleotide Polymorphisms，dbSNP)[8]，确定第22号染色体中SNPs所处的基因名称。结果表明，在第22号染色体上32 668个SNPs中，有15 617个SNPs处于428个已知基因内，据此将其划分成428个SNPs集合；其余17 051个SNPs未处于已知基因区域，将其划分成相应数量的仅含单个SNP的集合(即17 051个SNPs集合)。

其次，在每个SNPs集合中进行PCA，选择PC$_1$建立主成分logistic回归模型，对主成分的偏回归系数进行似然比检验得到*P*值，以检验该SNPs集合与疾病的关联。

最后，将上述步骤重复100次。计算每个SNPs集合重复100次假设检验所得到的*P*值的中位数。第22号染色体的全基因扫描结果见图1。并对基因按照*P*值进行排序，其中*P*值最小的5个基因的结果见表1。

基于基因水平的主成分logistic回归模型分析结果表明，在所有基因假设检验的*P*值中，TNRC6B基因的*P*值为最小：2.42×10^{-5}，采用Bonferroni校正后差异有统计学意义，结果提示该基因区域可能为致病位点所在区域。这与事先模拟试验的参数设定相符。模拟研究表明采用基于基因水平的主成分logistic回归模型，能够检测基因组中与疾病有关联的基因区域，且具有一定检验效能。

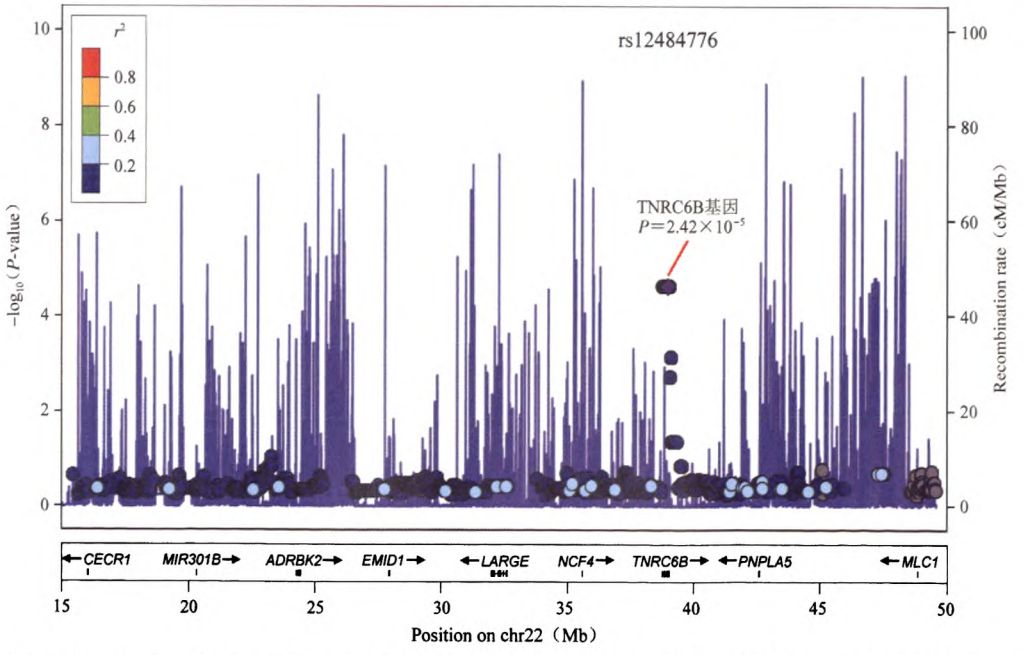
表1　基于基因水平的主成分logistic回归模型分析GWAS基因型模拟数据(第22号染色体)基因*P*值排序

基因名称	染色体位置(Mb)	*P*值[a]			
		中位数	最小值	均数	最大值
TNRC6B	38.77 ~ 39.06	0.000 024 2[b]	1.04e-13	0.003 662 9	0.129 108 0
SGSM3	39.08 ~ 39.12	0.000 733 0	1.30e-09	0.052 172 4	0.826 181 2
ADSL	40.74 ~ 40.76	0.001 940 5	8.61e-10	0.067 858 1	0.755 545 1
MKL1	40.81 ~ 41.03	0.043 481 4	7.62e-07	0.147 735 2	0.834 117 7
C22orf13	24.94 ~ 24.95	0.093 377 9	1.81e-06	0.195 636 5	0.927 412 1

注：按*P*值中位数排序，仅列*P*值最小的5个基因区域结果；[a]基于100次模拟计算*P*值的描述统计量；[b]采用Bonferroni校正后差异有统计学意义；模拟试验设定的致病位点为rs12484776，染色体物理位置为38.9828 Mb，位于TNRC6B基因区域

讨　论

本研究针对全基因组关联研究数据，通过利用生物学先验信息将SNPs数据降为具有生物学功能的基因水平，在每个基因中采用基于主成分的logistic回归模型来检测基因与疾病之间的统计学关联。模拟试验结果表明该方法能够检测遗传致病位点所在的区域，在分析GWAS资料时具有一定的检验效能。



注：左纵坐标中*P*值为100次模拟后*P*值的中位数；右纵坐标为重组率；横坐标为第22号染色体物理位置；模拟试验设定的致病位点为rs12484776，物理位置为38.9828 Mb，位于TNRC6B基因区域

图1　基于基因水平的主成分logistic回归模型分析GWAS基因型模拟数据(第22号染色体)基因组扫描结果

在GWAS中，采用基于主成分的logistic回归模型主要具有以下优点：

首先，它是一种多位点分析的方法。该方法通过主成分提取基因内多个位点遗传变异的综合信息，从基因水平上研究遗传变异与疾病的关联，这样的分析策略更加符合复杂性疾病的致病机制。许多研究已表明[9-11]，基于多个SNPs位点的分析要优于传统的基于单个SNPs分析策略。

其次，它克服了传统方法在分析GWAS数据时所遇到的多重共线性问题。在GWAS中，由于多个位点之间存在关联性即连锁不平衡，传统多变量回归模型将会由于多重共线性使得模型的估计变得不稳定。在主成分logistic回归模型中，由于模型中的各项都是互相独立的，因此在考虑SNPs之间相关性的同时又解决了共线性问题。

最后，采用基于基因水平的主成分回归方法，虽不能完全避免多重比较的问题，但是能够降低多重比较的次数。如果仅在某一个候选基因区域采用主成分回归的方法，此时不需要考虑多重比较校正问题。如果在全基因组范围内采用基于基因水平的主成分回归的方法，则需要考虑多重比较的校正问题。此时，相对于基于单个位点的分析策略，校正的次数将会大大减少。

但是，基于基因水平的主成分logistic回归模型也存在一些局限性：

首先，在GWAS中主成分的生物学意义有时较为难以解释。尤其对于GWAS数据或者基因表达数据，主成分通常是非常多SNPs或者基因的线性组合，导致主成分的意义不容易解释。此外，由于该方法是多位点的统计方法，结果仅提示某个基因区域有统计学意义，该方法本身并不能提示该区域内哪个或哪些SNPs与疾病有关联。为此，有研究者提出检验因子载荷的方法，即通过因子载荷检验其中对疾病影响最大的SNP位点，或者采用有监督的（supervised）或者稀疏的（sparse）主成分回归模型[12,13]，以提高对结果的解释。

其次是主成分个数的选择。回归模型中主成分个数越多，能够解释的变异就越多，但模型检验的自由度将会大大增加，检验效能将会降低。因此模型中主成分个数的选择要同时考虑主成分对总方差的解释比例，以及模型检验自由度这两个影响因素，在两者中选择恰当的平衡点。一些研究者提出了选择主成分个数的方法[14,15]，例如，选择能够解释50%～70%以上累计变异的主成分，或选择特征根＞1的主成分。在实际资料的分析中，如何选择合适的主成分来建模并无定律，应选择多个不同数量的主成分分别建模并比较，以选择既能符合生物学解释又能具有较好拟合效果的回归模型。

最后是SNPs集合的划分方法。本研究基于基因进行划分，由于许多SNPs所处的基因还并不明确，因此存在一定的局限性。为了能够覆盖全基因组数据，在实际工作中，还可结合其他方法进行划分，例如根据基因通路、单倍型域、基于基因区域的保守程度等进行划分。

参 考 文 献

[1] Moore JH, Ritchie MD. The challenges of Whole-Genome approaches to common diseases. JAMA, 2004, 291(13): 1642-1643.

[2] Ma S, Dai Y. Principal component analysis based methods in bioinformatics studies. Brief Bioinform, 2011, 12(6): 714-722.

[3] Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet, 2010, 86(6): 929-942.

[4] Marchini J, Howie B, Myers S, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet, 2007, 39(7): 906-913.

[5] Spencer CC, Su Z, Donnelly P, et al. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet, 2009, 5(5): e1000477.

[6] Ayers KL, Cordell HJ. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genet Epidemiol, 2010, 34(8): 879-891.

[7] Gao Q, He Y, Yuan Z, et al. Gene- or region-based association study via kernel principal component analysis. BMC Genet, 2011, 12: 75.

[8] Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res, 2001, 29(1): 308-311.

[9] Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. Am J Hum Genet, 2004, 75(3): 353-362.

[10] Buil A, Martinez-Perez A, Perera-Lluna A, et al. A new gene-based association test for genome-wide association studies. BMC Proc, 2009, Suppl 7: S130.

[11] Huang H, Chanda P, Alonso A, et al. Gene-based tests of association. PLoS Genet, 2011, 7(7): e1002177.

[12] Bair E, Hastie T, Dabashis P, et al. Prediction by supervised principal components. J Acous Soc Am, 2006, 102: 119-137.

[13] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. J Comp Graph Stat, 2006, 15: 262-286.

[14] Chen X, Wang L, Hu B, et al. Pathway-based analysis for genome-wide association studies using supervised principal components. Genet Epidemiol, 2010, 34(7): 716-724.

[15] Saris CG, Horvath S, van Vught PW, et al. Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. BMC Genomics, 2009, 10: 405.

（收稿日期：2011-12-07）
（本文编辑：张林东）