·基础理论与方法·

核函数logistic回归模型在全基因组关联研究中的应用

沃红梅 易洪刚 潘红星 唐少文 赵杨 陈峰

【导读】 探讨基于基因水平的核函数logistic回归模型及其在全基因组关联研究中的应用。以全基因组关联研究模拟数据为例，介绍核函数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，在所有已知基因检验结果中致病位点所在基因假设检验的P值最小。结果提示基于基因水平的核函数logistic回归模型能够充分提取和综合基因中多个遗传突变位点信息，降低统计学检验的自由度，同时还能够控制多种协变量因素和交互作用，在检测致病基因与疾病关联时具有一定的效能。

【关键词】 核函数；Logistic回归；全基因组关联研究

Application of gene-based logistic kernel-machine regression model on studies related to the genome-wide association

WO Hong-mei[1], *YI Hong-gang*[1], *PAN Hong-xing*[2], *TANG Shao-wen*[1], *ZHAO Yang*[1], *CHEN Feng*[1]. 1 *Department of Epidemiology and Biostatistics, School of Public Health, Nanjing Medical University, Nanjing* 211166, China; 2 *Jiangsu Provincial Center for Disease Control and Prevention*

Corresponding author: CHEN Feng, Email: fengchen@njmu.edu.cn

This work was supported by grants from the National Natural Science Foundation of China (No. 81202283, 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 logistic kernel-machine regression model and its application in genome-wide association study (GWAS). Using the simulated genome-wide single-nucleotide polymorphism (SNPs) genotypes data, we proposed a practical statistical analysis strategy-named 'the logistic kernel-machine regression model', based on the gene levels to assess the association between genetic variations and complex diseases. The results from simulation showed that the P value of genes in related diseases was the smallest among all the genes. The results of simulation indicated that not only it could borrow information from different SNPs that were grouped in genes and reducing the degree of freedom through hypothesis testing, but could also incorporate the covariate effects and the complex SNPs interactions. The gene-based logistic kernel-machine regression model seemed to have certain statistical power for testing the association between genetic genes and diseases in GWAS.

【Key words】 Kernel function; Logistic regression; Genome-wide association study

全基因组关联研究(GWAS)已成为人类复杂性疾病遗传易感性的主要研究策略。但GWAS也面临着许多统计学问题：①现有的生物学技术已经能够在一个基因芯片上同时检测上百万个位点的单核苷酸多态性(SNPs)，因此在GWAS数据中就会产生较大的变量数p(即SNPs)和较小的样本量n的问题(large p small n problem)，即"维度灾难"[1]。②目前GWAS的分析策略仍以基于单个SNP的分析为主[2]。然而该策略存在两方面问题。首先从统计学角度看，由于多重比较(multiple comparison)的次数多达几十万次，必须校正才能有效控制第一类错误(type Ⅰ error)[3]。实际工作中经常采用Bonferroni方法校正，但是研究表明这种方法非常保守[4]。其次从生物学角度看，复杂性疾病往往受多个不同位点、基因或者通路的影响，而不仅仅是单个位点突变而致病，因此单位点分析策略并不符合复杂性疾病的

DOI: 10.3760/cma.j.issn.0254-6450.2013.06.023

基金项目：国家自然科学基金(81202283, 81072389, 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：211166 南京医科大学公共卫生学院流行病与卫生统计学系(沃红梅、易洪刚、唐少文、赵杨、陈峰)；江苏省疾病预防控制中心(潘红星)

通信作者：陈峰，Email: fengchen@njmu.edu.cn

致病机制。因此，在GWAS中大量的假设检验经过多重比较调整后，再采用传统的单位点分析策略时，很有可能会错失那些较弱的、但真正与疾病相关联的位点。由此可见，基于单个SNP的分析策略显然不能满足实际需要。因此，许多研究者提出各种GWAS的分析策略和统计方法[5,6]。其中Kwee等[7]、Wu等[8]提出一种采用非参数核函数(kernel function)的方式建立基因型数据和疾病表型之间非线性关系的回归模型，由于具有能够利用生物学先验信息及具有较多的统计学特性等优点，开始在GWAS中得到应用并取得一定效果。本研究主要介绍基于基因水平的核函数logistic回归模型在GWAS中的建模策略和应用。

基本原理

在分析GWAS资料时，首先基于生物学先验信息，将全基因组中的SNPs划分成具有生物学特征的SNPs集合(SNPs set)，然后在每个具有生物学意义的SNPs集合中，建立核函数logistic回归模型，并采用方差成分得分检验(variance-component score test)计算每个SNPs集合的P值，以评价同一SNPs集合中所有SNPs的联合效应与疾病的关联。

具体方法和分析步骤：

1. 划分SNPs集合：根据生物学先验信息，对全基因组中SNPs进行划分，形成有生物学意义的SNPs集合，例如基于基因、基因通路等。根据基因划分是最常用的一种策略，将位于同一个基因中(或者相邻区域)的SNPs划分成一个SNPs集合。

2. 建立核函数logistic回归模型：假设采用以人群为基础的病例对照设计，对n个独立的个体进行基因型分型。设y_i为第i个个体表型的数据，$y_i=1$表示病例，$y_i=0$表示对照。在每个SNPs集合中，包含p个SNPs，设$Z_{i1}, Z_{i2}, \cdots, Z_{ip}$为第$i$个个体($i=1,2,\cdots,n$)的基因型数据。采用相加模式，$Z_{ij}=0,1,2$分别表示第$i$个个体第$j$个SNPs位点的最小等位基因(minor allele)个数。设$x_{i1}, x_{i2}, \cdots, x_{im}$为$m$个用于调整的协变量，例如人口学变量、环境因素变量等。

为检验在调整了其他协变量情况下，包含p个SNPs的集合是否与疾病存在关联，对于第i个个体，建立核函数logistic回归模型

$$\text{logit}\,P(y_i=1)=\alpha_0+\alpha_1x_{i1}+\alpha_2x_{i2}+\alpha_mx_{im}+h(Z_{i1},Z_{i2},\cdots,Z_{ip}) \quad (1)$$

式中，α_0是常数项，$\alpha_1,\alpha_2,\cdots,\alpha_m$为协变量的偏回归系数。核函数$h(Z_{i1},Z_{i2},\cdots,Z_{ip})$为任意形式、未知的函数，是模型中最主要的部分。$h(\cdot)$综合了该集合中所有SNPs的遗传信息，表达了SNPs集合和疾病之间复杂的非线性关系。经数学证明[9]，$h(Z_{i1},Z_{i2},\cdots,Z_{ip})=h(\mathbf{Z}_i)=\sum_{i'=1}^{n}\gamma_{i'}K(\mathbf{Z}_i,\mathbf{Z}_{i'})$。可见，其中$h(\cdot)$完全由正半定核函数矩阵$K(\cdot,\cdot)$来决定。

3. 选择核函数：核函数矩阵$K(\cdot,\cdot)$决定了非参数函数$h(\cdot)$的形式，从而进一步影响了SNPs集合和表型之间的关联性。在遗传学中，用来衡量第i个个体和第i'个个体基因型相似性的一种常用方法就是计算该对个体状态同一(identical-by-state, IBS)的等位基因个数。因此本研究考虑采用IBS核函数：$K(\mathbf{Z}_i,\mathbf{Z}_{i'})=\sum_{j=1}^{p}\{2I(Z_{ij}=Z_{i'j})+I(|Z_{ij}-Z_{i'j}|=1)\}/2p$表达个体之间遗传相似性。

4. 方差成分得分检验：考虑到第i个个体患病概率只取决于$h(\mathbf{Z}_i)$。因此，为了检验SNPs集合与疾病关联，建立假设检验

$$H_0: \mathbf{h}(\mathbf{Z})=0 \quad (2)$$

为检验该假设，Liu等[9]证明$\mathbf{h}=\mathbf{K}\gamma$。式中，K为$n\times n$的IBS核函数矩阵$K(\mathbf{Z}_i,\mathbf{Z}_{i'})$；$\mathbf{h}=[h_1,h_2,\cdots,h_n]'$为每一个个体随机效应，服从一个均数为0方差为$\tau\mathbf{K}$的任意$F$分布。其中$\tau$体现了SNPs集合的遗传效应。则式(2)中对SNPs集合中遗传效应的假设检验等价于式(3)中检验参数τ是否为0，即

$$H_0: \mathbf{h}(\mathbf{Z})=0 \Leftrightarrow H_0: \tau=0 \quad (3)$$

因此，可采用通过方差成分得分法以检验参数τ是否为0，即

$$Q=\frac{(y-\hat{\mathbf{p}}_0)'\,\mathbf{K}\,(y-\hat{\mathbf{p}}_0)}{2} \quad (4)$$

式中，$\text{logit}\,\hat{p}_{0i}=\hat{\alpha}_0+\hat{\alpha}_1x_{i1}+\hat{\alpha}_2x_{i2}+\cdots+\hat{\alpha}_mx_{im}$，常数项$\hat{\alpha}_0$和协变量偏回归系数$\hat{\alpha}_j$基于零模型得到估计。最后基于$Q$所服从的尺度参数为$k$以及自由度为$\nu$的$\chi^2$分布，获得假设检验中的$P$值[8]。

实例分析

本研究采用病例对照设计，使用HapGen软件[10]模拟产生人类基因组GWAS基因型数据，采用核函数logistic回归模型进行分析。以第22号染色体基于基因分析的结果为例，介绍建模策略和结果解释。

本研究采用文献[11,12]的方法模拟产生1000个病例，1000个无关联对照的全基因组基因型数据。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，采用HapGen模拟产生第22号染色体物理位

置为14.43～49.58 Mb的基因型数据，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，MAF）设为15%。致病等位基因杂合子致病的优势比（*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）[13]，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个已知基因中，据此将其划分成428个SNPs集合（表1），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，在每一个SNPs集合中，进行IBS核函数logistic回归模型分析，采用方差成分得分检验得到每个分析集合的*P*值，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2。处于基因区域的第22号染色体的基因组扫描结果见图1。

最后，基于基因水平的核函数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，结果提示该基因区域为致病位点所在区域。这与事先模拟试验的参数设定相符。与传统的基于单个位点的logistic回归模型结果相比较，核函数logistic回归模型结果中致病位点的*P*值更小且非致病位点（噪声位点）的*P*值更大，因此能够更容易地从基因组中大量非致病位点（噪声位点）中筛选出与疾病关联的位点。模拟研究表明采用基于基因水平的核函数logistic回归模型，

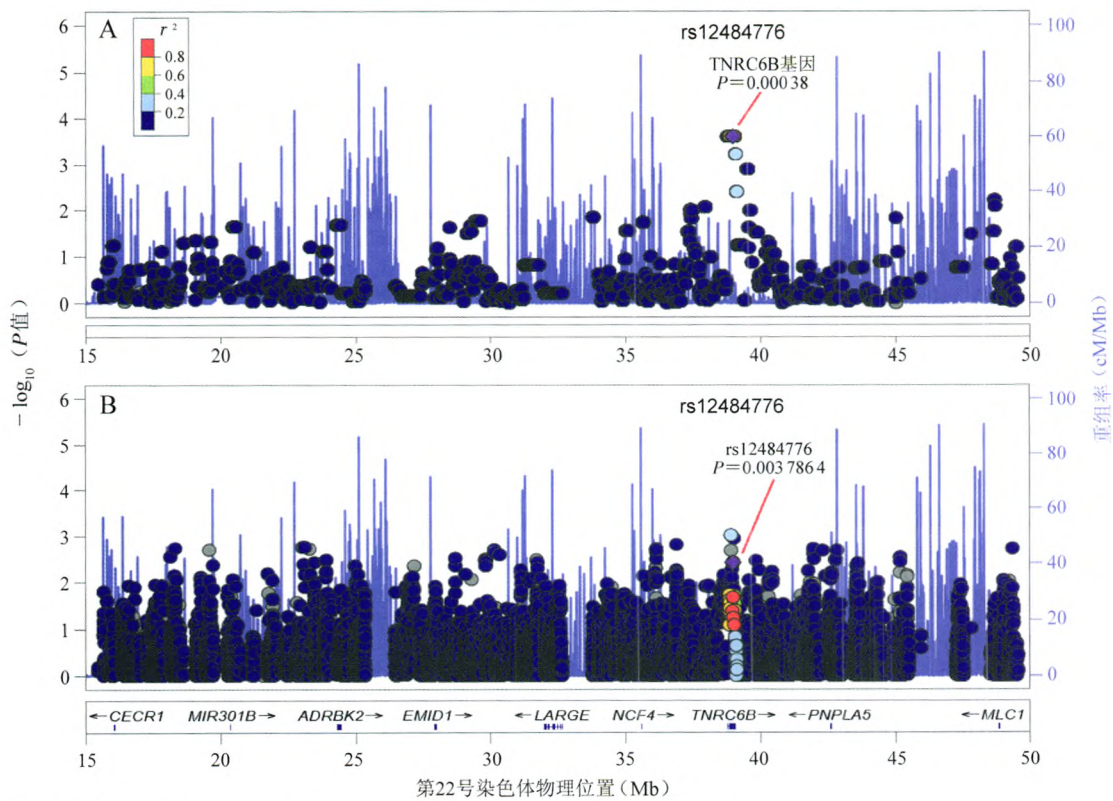
表1 SNPs集合中SNPs的频数分布

SNPs集合中SNPs个数	SNPs集合数	构成比(%)	SNPs集合中SNPs个数	SNPs集合数	构成比(%)
1～10	160	37.38	41～50	26	6.07
11～20	81	18.93	51～100	45	10.51
21～30	51	11.92	≥101	33	7.71
31～40	32	7.48	合计	428	100.00

表2 基于基因水平的核函数logistic回归模型分析GWAS基因型模拟数据（第22号染色体）基因*P*值排序结果

序号	基因名称	染色体位置(Mb)	*P*值
1	TNRC6B	38.77～39.06	0.000 38
2	ADSL	40.74～40.76	0.000 75
3	SLC25A17	39.50～39.55	0.001 62
4	SGSM3	39.08～39.12	0.005 95
5	XPNPEP3	39.58～39.65	0.009 55

注：仅列出*P*值最小的5个基因区域结果；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



注：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 基于基因水平的核函数logistic回归模型（A）和传统的logistic回归模型（B）分析GWAS基因型模拟数据（第22号染色体）基因组扫描结果

能够检测基因组中与疾病有关联的基因区域，并具有一定的检验效能。

讨　论

基于基因水平的核函数logistic回归模型是一种能够把生物学先验信息和统计分析方法充分结合的回归模型。其分析策略是首先利用各种生物学先验信息降维，在具有生物学意义的分析单位上（如基于基因水平）进行分析；其次采用核函数体现SNPs集合中多个位点的遗传效应，并检验SNPs集合与疾病之间复杂的非线性关系。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。

在GWAS中，基于基因水平的核函数logistic回归模型主要具有以下优点：

1. 该模型可以充分利用生物学先验信息。由于复杂性疾病致病机制往往受到多个不同基因或者基因通路的影响，而不仅仅是单个位点突变而致病，因此传统单位点分析显然不能满足实际需要。本研究在GWAS数据分析中基于先验生物学信息，从传统的单位点分析，转变为以基因为基础的分析。该分析策略可以综合、汇聚同一基因或者通路中多个SNPs较弱的遗传变异信息，更符合生物学致病机制及更接近真正的致病过程，具有生物学解释的优势；此外还能够降低多重比较次数，以提高模型估计效果和检验效能。

2. 核函数logistic回归模型在统计学上具有诸多优良的方法学特性。首先，采用非参数核函数是该模型一个重要特点。核函数矩阵$K(\mathbf{Z}_i, \mathbf{Z}_{i'})$是针对SNPs集合中所有个体，测量任意2个个体（第i个个体和第i'个个体）基因型信息相似性的函数，能够综合体现同一个SNPs集合中多个SNPs位点的遗传效应。其次，该模型能够建立遗传效应和疾病之间非线性关系。与传统广义线性回归模型需要线性假设不同的是，核函数矩阵将基因型信息从原始空间投影到另外一个空间，然后在新空间中建立$h(\cdot)$与疾病表型的广义线性模型。因此，该模型能够检验遗传效应与因变量之间复杂的非线性关系，而非传统方法所假设的遗传与疾病之间为线性关系。最后，该模型能够利用多个位点间的相关性即连锁不平衡（linkage disequilibrium, LD）以提高检验效能。传统单位点分析均需要位点间互相独立的假设。然而，基因组中各个位点间存在着不同程度LD。在核函数回归模型的方差成分得分检验中，自由度ν的估计值将会随着SNPs之间相关性的增加而降低。这就提示该模型能够利用LD，通过自由度的自适应估计来获得更高的检验效能。

基于基因水平的核函数logistic回归模型也存在局限性。①本研究采用基于已知基因的方法具有一定局限性，因为许多SNPs所处的基因还并不明确。除本研究外还可采用基于基因通路、单倍型域、基因区域保守程度等方法。实际工作中综合使用各种划分策略，则能够提供全基因组划分的覆盖。②在遗传学关联研究中，有多种核函数矩阵可供选择，例如线性核函数（linear kernel）、高斯核函数（Gaussian kernel）以及加权的IBS（weighted-IBS kernel）核函数等。尽管许多研究者对多种核函数矩阵的效果进行了比较研究，提供了不同的选择策略，但是在GWAS中的应用效果和效率仍需进一步评价[8,14,15]。

参 考 文 献

[1] Kelemen A, Vasilakos AV, Liang Y. Computational intelligence in bioinformatics: SNP/haplotype data in genetic association study for common diseases. IEEE Trans Inf Technol Biomed, 2009, 13(5): 841-847.

[2] Peng G, Luo L, Siu H, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet, 2010, 18(1): 111-117.

[3] Johnson RC, Nelson GW, Troyer JL, et al. Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics, 2010, 11: 724.

[4] Shi Q, Pavey ES, Carter RE. Bonferroni-based correction factor for multiple, correlated endpoints. Pharm Stat, 2012, 11(4): 300-309.

[5] Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. Bioinformatics, 2010, 26(4): 445-455.

[6] Yaspan BL, Veatch OJ. Strategies for pathway analysis from GWAS data. Curr Protoc Hum Genet, 2011, Chapter 1: Unit1 20.

[7] Kwee LC, Liu D, Lin X, et al. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet, 2008, 82(2): 386-397.

[8] 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.

[9] Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics, 2008, 9: 292.

[10] 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.

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

[12] Yi HG, Wo HM, Zhao Y, et al. Gene-based principal component logistic regression model and its application on genome-wide association study. Chin J Epidemiol, 2012, 33(6): 622-625. (in Chinese)
易洪刚，沃红梅，赵杨，等. 基于基因水平的主成分logistic回归模型在全基因组关联研究中的应用. 中华流行病学杂志，2012，33(6)：622-625.

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

[14] Lin WY, Schaid DJ. Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. Genet Epidemiol, 2009, 33(3): 183-197.

[15] Mukhopadhyay I, Feingold E, Weeks DE, et al. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. Genet Epidemiol, 2010, 34(3): 213-221.

（收稿日期：2012-10-31）
（本文编辑：张林东）