·学习·发现·交流·

全基因组关联研究中的统计分析方法

陈峰　柏建岭　赵杨　荀鹏程

【导读】 随着人类基因组计划的完成，疾病的全基因组关联研究成为可能。该类研究的数据特点是：高维、小样本。面对浩瀚的数据，传统分析方法受到严重挑战。文中介绍全基因组关联研究中的数据分析策略和步骤，包括质量控制、分析、结果表示等，并对全基因组关联研究的局限性和目前统计分析方法的不足进行讨论。

【关键词】 全基因组关联研究；质量控制；数据管理；统计分析

Statistical methodologies used in genome-wide association studies

CHEN Feng, BAI Jian-ling, ZHAO Yang, XUN Peng-cheng. Department of Epidemiology and Health Statistics, 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. 30901232, 81072389) *and Major Program of the Natural Science Foundation of the Jiangsu Higher Education Institutions of China* (No. 10KJA33034).

【Introduction】 In lieu of large samples of cases and/or controls with hundreds of markers spreading throughout the human genome, researchers started to notice the dramatic increase of genome-wide association study (GWAS) for complex disorders, in the last 5 years. This paper highlights the statistical challenges in such huge-scale genetic studies, and introduces the analytical strategies and steps for handling GWAS data. Such issues as quality control of data, population stratification, methods available to data analysis and results presentation, replication, as well as the limitations of GWAS studies and the challenges presenting for statistics, are addressed.

【Key words】 Genome-wide association study; Quality control; Data management; Statistical analysis

全基因组关联研究(GWAS)是利用全基因组高通量测序技术，对研究对象的基因组中序列变异或单核苷酸多态(SNPs)进行分型，并利用生物统计学和生物信息学的方法，检验SNP或基因与复杂疾病或可测性状(observable trait)的关联性研究方法。该项技术能为确定疾病发病易感位点、区域和相关基因，寻找疾病的标记物，为阐明疾病的遗传机制，进而为疾病的早期诊断、个性化治疗、新药研发和特异性防治措施提供依据[1]。

自从2005年*Science*上发表第一篇GWAS的论文后，GWAS如雨后春笋一样不断涌现，逐年增加。根据http://www.genome.gov/和PubMed上的资料，截至2010年10月17日，公开发表的GWAS有702篇，其中我国科学家公开发表了10篇[2-11]。

DOI: 10.3760/cma.j.issn.0254-6450.2011.04.018

基金项目：国家自然科学基金(30901232, 81072389)；江苏省高校自然科学基金重大项目(10KJA33034)

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

GWAS主要采用病例对照、随访和以家庭成员为基础的家系研究。在已经发表的702篇GWAS中，病例对照研究占238篇。GWAS区别于候选基因研究(candidate gene association study)，前者是探索性的(non-hypothesis-driven)，后者是在得到一些生物学或流行病学线索后对选定的基因/位点进行的研究。因此GWAS所得到的结果更需要后续的验证(validation)。GWAS也区别于基因表达研究(gene expression study)，前者的SNP变量为二分类或三分类，而后者中的相应变量往往是连续性的。SNP是遗传变异中最常见的一种，GWAS中往往考虑SNP，也有考虑拷贝数变异的(copy number variation, CNV)。

GWAS的不断发展，推动了相应的统计分析方法的研究和专门统计软件的开发，例如PLINK、GenABEL等[12,13]。为此本文论述GWAS数据统计学分析的前期处理、分析策略以及结果展示等。

1. 数据清理：GWAS的首要问题是数据清理(data cleaning)，即质量控制。与候选基因研究、临

床随访研究不同，GWAS涉及数十万乃至上百万个变量，数据量庞大，需要更严格的质量控制标准。数据清理分两个部分，一是针对变量（即SNP），另一是针对个体（即样品）。前者需要考虑分型率（genotyping calling rate）、次等位基因频率（minor allele frequency，MAF）和Hardy-Weinberger平衡（HWE）等；后者主要考虑缺失率（missingness rate）、人群分层（population stratification，PS）和种族混杂，研究个体间的独立性（independency）以及性别核查（sex check）等问题。

（1）分型率和准确性：几乎所有GWAS的分型是由Affymetrix（Santa Clara，California，USA）和Illumina（San Diego，California，UAS）两大公司完成。分型准确性的评价可以有两种方法。一是对同一个体的2份样品的结果进行比对，二是对HapMap中LD $r^2=1$的位点，比较它们在GWAS数据中的关系[14]。分型率是反映基因分型质量的重要指标，常以95%、98%或99%作为界值，剔除分型率低的SNP。分型准确性和分型率在基因公司内部也是重要的质量控制步骤。

（2）次等位基因频率：GWAS只考虑多态性（polymorphisms）位点，不考虑没有提供与疾病关联信息的无变异单态性（monomorphism）位点。而每个位点上有2个等位基因（allele），在不同人群中，2个等位基因的频率不同。频率高的等位基因称为主要等位基因（major allele），反之为次等位基因（minor allele）。

GWAS中，如果MAF很低时，一方面说明变异小，提供的与疾病关联信息少；另一方面，关联性检验的统计学效能很低。例如，在病例对照研究中，同样是1000个病例和1000个对照，MAF分别是0.01和0.20时，检出理论$OR=1.5$的把握度分别为16.8%和97.0%。可见，MAF对检验效能有比较大的影响。因此，GWAS中剔除MAF较低的SNP，目前剔除MAF的界值常选为0.01～0.05。

（3）HWE：这是群体遗传中的重要法则，在没有进化影响下当基因一代一代传递时，群体的基因频率和基因型频率将保持不变，两者的关系也保持不变，且前者可以确定后者。不满足HWE的群体，说明可能存在近亲婚配、遗传漂移、严重突变、人群分层等，代表性差，不能作进一步分析。由于疾病的发生可能导致遗传不平衡，因此在病例对照研究中HWE检验只针对对照组。GWAS中HWE检验水准常取10^{-4}～10^{-6}。

（4）缺失率：个体SNP的缺失率是反映DNA样本质量的重要指标，如果缺失多，则说明该个体的DNA样品质量差。这一步通常在基因公司内部也是重要的质量控制步骤。常用0.01、0.02或0.05作为界值，剔除缺失率大于界值的个体。

必要时，对同一SNP在病例组和对照组的缺失率进行比较，以判断是否为随机缺失。

（5）地域差异和人群分层：在病例对照研究、随访研究中，研究样本的地域差异（geographical variation）和人群分层（population stratification，PS）是种族混杂（confounding by ethnicity）的表现，将导致虚假关联。因此，GWAS中需要阐述是否存在PS，若存在，如何校正，并说明校正后的效果。

PS的判别常用膨胀系数（genomic control inflation factor，λ_{GC}）表示，它是所有SNP检验统计量（例如Cochran-Armitage趋势检验的χ^2值）的中位数（或均数）与理论分布中位数（或均数）的比值[15,16]。$\lambda_{GC}=1$表示没有人群分层，$\lambda_{GC}>1$表示有人群分层。QQ图（quantile-quantile plot）可以帮助判断。当存在PS，且可以获得个体种族的地域信息时，应该根据种族或地域信息进行子集分析。事实上，实际工作中很难确切知道个体种族或地域信息，故常采用GC（genomic control）法[15]、SA（structured association）法[17]或主成分法。

Price等[18]建议用Eigenstrat法检测校正PS。该法是利用相对独立的SNP（LD r^2较小，例如0.05）估计主成分，在关联性模型中增加主成分作为协变量，从而达到校正PS的目的。至于采用多少个主成分进行校正，主要看λ_{GC}的变化，也可用以采用Tracy-Wisdom方法进行检验。该方法可以用Price等[18]在Linux系统下开发的EIGENSTRAT软件实施。

调查资料结合HapMap中不同人种的资料，利用Eigenstrat法可以检测种族异常值，例如检出可能的混血儿。人群分层不仅仅存在于不同的种族之间，在同一人种中也同样存在此问题。如祖先同为来自欧洲的高加索人（Caucasian），基因型频率存在南北差异（north-south gradient）[19,20]。相比而言，我国汉族人群分层现象虽然不是很突出，但有研究表明[21,22]，汉族人的基因型频率也同样存在着南北梯度。当考虑少数民族时，PS就更明显了。

如果样本存在人群分层，不校正则将增加假阳性；而不必要的校正常导致检验效能降低。因此，应正确判断是否需要校正，以及如何校正。

避免PS的一个有效方法是采用家系研究。

(6)个体间的独立性:无论是病例对照研究还是随访研究,都需要满足有一个统计学假设,即研究个体间是相互独立的(independent)。如果研究个体间不独立,例如研究样本中包含了有血缘关系的一、二级亲属,则分析时需要考虑剔除这些非独立的个体。

个体间是否有血缘关系可以用同源(identical-by-descent, IBD)等位基因的概率分布来判断。例如,兄弟对的同源等位基因个数的概率分布为:$P(IBD=0)=0.25$, $P(IBD=1)=0.5$, $P(IBD=2)=0.25$;第一代单表兄弟的同源等位基因个数的概率分布为:$P(IBD=0)=0.75$, $P(IBD=1)=0.25$, $P(IBD=2)=0$。如果个体是独立的,则2个个体的$P(IBD=0)=1$。

利用该方法,还可以判断重复的个体,送检样本相互污染的个体等。例如,笔者在分析哈佛大学肺癌全基因组资料时,发现2个基因信息完全相同的样品,经进一步核对,确认为同一个人的2份样品,一份是首次诊断肺癌时被纳入研究的,另外一份是3年后复发时再次被纳入研究的。分析时,剔除了后者。

(7)性别核查(sex check):是基于各SNP在X染色体上的杂合率(heterozygosity rates)来进行的。如果调查表中报告的性别与基于X染色体估计的性别不一致,则需要进一步复核。笔者认为,当报告性别与估计性别不一致时,如果所研究的疾病或性状与性别关系不大,则可以考虑用估计的性别替代(impute)报告性别;如果有关,则需剔除。例如,在Alzheimer的GWAS中剔除了21个报告性别与估计性别不一致的研究对象[23];而在老年听力障碍的GWAS中用估计的性别代替报告性别[24]。

2. 关联研究:经过严格的质量控制后,就可以对清理后的数据进行关联性分析。关联性分析中需要考虑的统计学问题:生物学模式与统计学模型的选择、协变量调整和多重比较等问题。

(1)生物学模式与统计学模型的选择:生物学模式包括可加模式(additive model, trend model)、显性模式(dominant model)、隐性模式(recessive model)和共显性模式(co-dominant model)。如某个位点上的2个等位基因分别为A和a,则该位点有3种基因型(AA, Aa, aa)。不妨设A为野生等位基因,a为变异等位基因。则AA称为野生纯合型(common homozygous),aa称为变异纯合型(variant homozygous),Aa称为杂合型(heterozygous)。

在拟合关联模型,SNP于不同的生物学模式时,进入模型的形式不同。

可加模式中SNP以等级形式进入模型;显性模式和隐性模式中SNP以二分类形式进入模型,自由度为1;共显性模式中,SNP以哑变量的形式进入模型,自由度为2。即

相加模式:aa=2 *vs.* Aa=1 *vs.* AA=0

显性模式:(Aa或aa)=1 *vs.* AA=0

隐性模式:aa=1 *vs.* (Aa或AA)=0

共显性模式:aa *vs.* AA, Aa *vs.* AA

另外一种模式称为等位基因模式(allelic model),即a与A的比较。由于一个位点有2个等位基因,所以在Allelic模式中,样本含量是加倍的。

在病例对照研究中,可加模式常用Cochran-Armitage趋势检验,其他模式常用χ^2或Fisher确切概率检验。当需要校正协变量和PS时,用logistic回归模型。在随访研究中,用生存分析(如Cox)模型;当结果变量为数量性状时,用多重线性回归模型。

(2)协变量调整:由于GWAS是探索性研究,因此其统计学模型仅仅考虑最基本的(年龄、性别等)、公认的协变量校正,而不去深入探讨协变量的选择。在有些研究中甚至不校正任何协变量。但在研究设计时,需要考虑病例组和对照组的均衡性,以控制偏倚,提高可比性。

(3)多重比较(multiple comparison):这是GWAS中比较突出的问题。事实上,当每一次假设检验控制Ⅰ类错误为α时,K次独立的假设检验总的Ⅰ类错误就是$1-(1-\alpha)^K$,如果要将总的Ⅰ类错误率控制在0.05,则每一个SNP检验的水准就要控制得非常小。

控制Ⅰ类错误的方法很多。常用也是最方便的方法是Bonferroni法。如GWAS中检验50万个SNP,则按照Bonferroni控制,检验水准为0.05/500 000=10^{-7}。这一方法过于保守,因为该方法假设所有的SNP间是独立的。事实上,GWAS中这一假设不成立。因此有人提出采用较宽松的检验水准,例如10^{-6}, 5×10^{-6}, 10^{-5}等[25]。此外也常用Sidak法控制假发现率(false discovery rate)。

Moskvina和Schmidt[26]根据SNP间相关系数的特征根进行校正,并提出一个快速的近似算法。根据其研究结果,对上述问题,检验水准约为0.05/(5 000 000×0.65)=1.5×10^{-7}。但目前应用不多。

对于有些疾病,单个位点的效应不是很强。在样本含量不是很大时(很大程度上处决于研究经费),检验效能往往很低。如果再考虑多重比较,则

几乎没有一个SNP有统计学意义。因此，很多GWAS在探索阶段并不考虑多重比较的问题，而是选择P值最小的几个SNP进入下一步的验证。甚至可以不计算假设检验的P值，而直接利用一些统计量对SNP进行排序，如用Bayes因子等[27]。

(4)基于基因和通路的关联研究：前述的方法都是针对单个SNP进行的分析。但SNP不是独立的，疾病发生也不仅仅是某一个位点的单独作用。为此基于基因(gene-based)、基于通路(pathway-based)的分析方法应运而生[28-32]。其主要思想是降维(dimension reduction)，将多个SNP的变异或关联性检验的P值用一个综合指标(如主成分、综合评分等)来表示，达到综合评价的目的。

(5)结论的图形表达：由于分析的SNP很多，GWAS的结果常用图形的方式表示。如图1是按染色体的顺序排列，以各染色体中各位点的碱基对的位置为横轴，以各SNP与疾病关联性检验的P值之变换值$-\log_{10}(P)$为纵轴的一个散点图，称为Manhattan plot[33]。而对P值取对数变换的目的是为了突出小的P值。例如，纵轴>7，表示$P<10^{-7}$。

在高维数据分析中，常常用图形表示数据的特征和结构，使结果一目了然[34]。

3. 验证研究：初期的GWAS是探索性研究，为了验证GWAS的发现，控制假阳性，常常采用多阶段研究(multistage procedure)，即在第一阶段的GWAS后，根据研究的结果，在另外一个或几个独立的研究样本中对阳性结果进行验证(validation phase)。例如，在乳腺癌的GWAS中，有人采用了三阶段研究设计[35]；在结直肠癌的GWAS中采用四阶段研究设计[36]。

验证阶段一般是在与探索阶段相一致的人群中进行，也可以同时包含不同人群或不同人种；验证可以是内部验证(internal validation)，也可是外部验证(external validation)；可以借助于已有的同类GWAS资料进行验证，也可以是针对小规模，或在某一DNA片段中更高密度的分型。后者更为常见。显然，多阶段研究既能有效控制假阳性结果，提高检验效能，又可以降低研究成本。2007年美国国立癌症研究所(NCI)和国立人类基因研究所(NHGRI)就将验证研究定为GWAS中必不可少的一部分[37]。

多阶段研究的结果可以采用分层分析或Meta分析的方法进行综合，这样既提高了把握度(power)，又可以控制不同阶段研究间的异质性(heterogeneity)。

4. GWAS的局限性：疾病的发生、发展及预后是复杂的。一种疾病所涉及的基因绝不是一种，而是多种基因和功能综合改变的结果，并又有环境因素的共同参与。基因变异也多种多样，且这种改变不是静态的，在疾病发生、发展过程中，基因在表达、功能、代谢等方面的变化也是动态变化的。因此，基于SNP的GWAS只是探索疾病机制的方法之一和一个重要环节[38]。

由于花费巨大，样本含量有限，基因效应微弱，再加上人群分层现象的存在及多重比较的控制要求等，GWAS并没有原设想的那样取得举世瞩目的成果，虽有新发现，更多的是验证了先前候选基因研究的结果。此外另一个主要原因是数据分析手段落后于实际需求。目前的GWAS中所采用的统计分析方法，仍然以传统的方法为主。由于致病机制的复杂性，基于单个SNP的分析显然不能满足实际需要。基于基因、基于通路的分析，基因-环境交互作用、基因-基因交互作用的分析，特别是高阶相互作用、交互作用的分析是今后研究的重点。但是，面对几十

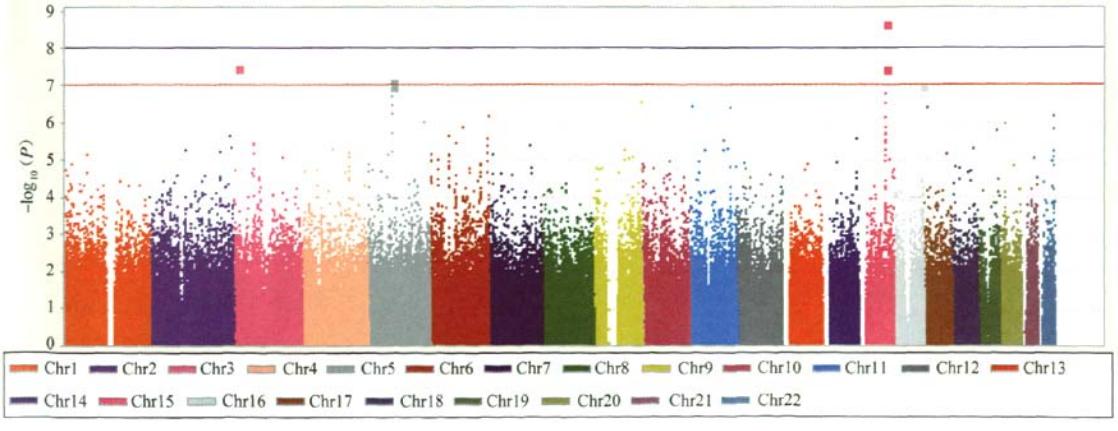

图1 某肺癌的全基因组病例对照研究模拟资料的Manhattan图

万个变量，目前的计算机系统很难胜任高阶交互作用的探索（例如，50万个SNP，即使是一阶的SNP-SNP交互作用，有超过10^{11}种组合）。因此，GWAS数据的深入分析和挖掘亟需新的分析思路、策略和手段。

参 考 文 献

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

[2] Zhang XJ, Huang W, Yang S, et al. Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21. Nat Genet, 2009, 41(2): 205-210.

[3] Han JW, Zheng HF, Cui Y, et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat Genet, 2009, 41(11): 1234-1237.

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

[5] Lei SF, Yang TL, Tan LJ, et al. Genome-wide association scan for stature in Chinese: evidence for ethnic specific loci. Hum Genet, 2009, 125(1): 1-9.

[6] Guo Y, Tan LJ, Lei SF, et al. Genome-wide association study identifies ALDH7A1 as a novel susceptibility gene for osteoporosis. PLoS Genet, 2010, 6(1): e1000806.

[7] Bei JX, Li Y, Jia WH, et al. A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. Nat Genet, 2010, 42(7): 599-603.

[8] Wu C, Xu B, Yuan P, et al. Genome-wide examination of genetic variants associated with response to platinum-based chemotherapy in patients with small-cell lung cancer. Pharmacogenet Genomics, 2010, 20(6): 389-395.

[9] Quan C, Ren YQ, Xiang LH, et al. Genome-wide association study for vitiligo identifies susceptibility loci at 6q27 and the MHC. Nat Genet, 2010, 42(7): 614-618.

[10] Zhang H, Zhai Y, Hu Z, et al. Genome-wide association study identifies 1p36.22 as a new susceptibility locus for hepatocellular carcinoma in chronic hepatitis B virus carriers. Nat Genet, 2010, 42(9): 755-758.

[11] Wang LD, Zhou FY, Li XM, et al. Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies susceptibility loci at PLCE1 and C20orf54. Nat Genet, 2010, 42(9): 759-763.

[12] Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet, 2007, 81(3): 559-575.

[13] Aulchenko YS, Ripke S, Isaacs A, et al. GenABEL: an R library for genome-wide association analysis. Bioinformatics, 2007, 23(10): 1294-1296.

[14] Teo YY. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. Curr Opin Lipidol, 2008, 19(2): 133-143.

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

[16] Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol, 2001, 20(1): 4-16.

[17] Pritchard JK, Stephens M, Rosenberg NA, et al. Association mapping in structured populations. Am J Hum Genet, 2000, 67(1): 170-181.

[18] Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet, 2006, 38(8): 904-909.

[19] Novembre J, Johnson T, Bryc K, et al. Genes mirror geography within Europe. Nature, 2008, 456(7218): 98-101.

[20] Campbell CD, Ogburn EL, Lunetta KL, et al. Demonstrating stratification in a European American population. Nat Genet, 2005, 37(8): 868-872.

[21] Xu S, Yin X, Li S, et al. Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am J Hum Genet, 2009, 85(6): 762-774.

[22] Chen J, Zheng H, Bei JX, et al. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am J Hum Genet, 2009, 85(6): 775-785.

[23] Carrasquillo MM, Zou F, Pankratz VS, et al. Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. Nat Genet, 2009, 41(2): 192-198.

[24] Huyghe JR, van Laer L, Hendrickx JJ, et al. Genome-wide SNP-based linkage scan identifies a locus on 8q24 for an age-related hearing impairment trait. Am J Hum Genet, 2008, 83(3): 401-407.

[25] Arking DE, Pfeufer A, Post W, et al. A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. Nat Genet, 2006, 38(6): 644-651.

[26] Moskvina V, Schmidt KM. On multiple-testing correction in genome-wide association studies. Genet Epidemiol, 2008, 32(6): 567-573.

[27] Wakefield J. Bayes factors for genome-wide association studies: comparison with *P*-values. Genet Epidemiol, 2009, 33(1): 79-86.

[28] Wang T, Elston RC. Improved power by use of a weighted score test for linkage disequilibrium mapping. Am J Hum Genet, 2007, 80(2): 353-360.

[29] Gauderman WJ, Murcray C, Gilliland F, et al. Testing association between disease and multiple SNPs in a candidate gene. Genet Epidemiol, 2007, 31(5): 383-395.

[30] Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet, 2007, 81(6): [Epub ahead of print]

[31] Wu MC, Zhang L, Wang Z, et al. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. Bioinformatics, 2009, 25(9): 1145-1151.

[32] Yu K, Li Q, Bergen AW, et al. Pathway analysis by adaptive combination of *P*-values. Genet Epidemiol, 2009, 33(8): 700-709.

[33] Barrett JC, Fry B, Maller J, et al. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics, 2005, 21(2): 263-265.

[34] Reimers M, Carey VJ. Bioconductor: an open source framework for bioinformatics and computational biology. Methods Enzymol, 2006, 411: 119-134.

[35] Thomas G, Jacobs KB, Kraft P, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet, 2009, 41(5): 579-584.

[36] Zanke BW, Greenwood CM, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet, 2007, 39(8): 989-994.

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

[38] Rodriguez-Murillo L, Greenberg DA. Genetic association analysis: a primer on how it works, its strengths and its weaknesses. Int J Androl, 2008, 31(6): 546-556.

（收稿日期：2010-10-29）
（本文编辑：张林东）