

综合报告

## 全基因组关联研究综述\*

潘东东

(云南大学数学与统计学院, 昆明, 650091)

李正帮

(华中师范大学数学与统计学院, 武汉, 430079)

张维 李启寨\*

(中国科学院数学与系统科学研究院, 北京, 100190)

### 摘要

本文是对近十年来科学前沿热点问题之一的全基因组关联研究(genome-wide association study, GWAS)的一个综述, 侧重于介绍其中所用到的统计分析方法, 讨论当前GWAS中存在的一些问题及挑战, 并就其发展前景作一个展望.

关键词: 病例-对照设计, 全基因组关联研究, 单核苷酸多态性, 稳健检验, 贝叶斯因子.

学科分类号: O213, Q348.

## §1. 引言

### 1.1 全基因组关联研究的定义

全基因组关联研究(genome-wide association study, 简记为GWAS)是一种用来寻找与人类复杂疾病或性状相关联的遗传变异的方法. 这里的遗传变异主要是指单核苷酸多态性(single nucleotide polymorphism, 简记为SNP). SNP是人类可遗传的变异中最常见的一种, 占有已知多态性的90%以上, 它在人类基因组中广泛存在, 平均每隔100至300个碱基对中就存在一个SNP, 其总数可达300万个或更多. GWAS将人类基因组中数以百万计的SNPs作为标记位点, 从目标人群中随机抽取一个较大规模的病例和对照样本利用基因芯片技术进行测序, 以获取各样本个体的DNA序列的观测数据并归整为各突变位点的基因型数据, 然后对全部SNPs位点逐个地进行统计检验及分析, 以期筛选出与复杂疾病有显著关联的遗传位点, 为多发性复杂疾病的诊断、预防和安全有效的治疗提供理论依据.

\*国家自然科学基金面上项目(11371353)、国家自然科学基金青年项目(11301465)、云南省应用基础研究计划青年项目(2013FD001)和云南大学校级基金项目(2012CG018)资助.

\*通讯作者, E-mail: liqz@amss.ac.cn.

本文2013年11月1日收到.

doi: 10.3969/j.issn.1001-4268.2014.01.008

## 1.2 全基因组关联研究与候选基因关联研究的联系和区别

候选基因关联研究(candidate gene association study, 简记为CGAS)与全基因组关联研究不同, 它只关注一部分事先指定好的基因中的一个或多个遗传变异与表现型或疾病状态之间的关联, 而GWAS则是在人类全基因组范围内扫描和检验遗传变异与复杂疾病的关联. CGAS和GWAS本质上都属于关联研究, 都是基于常见疾病/常见变异(common disease/common variant)假设, 即遗传因素对常见疾病易感性的影响体现在超过1%–5%的人群中出现一定数量的基因突变. 它们的工作原理也基本一致, 即通过比较SNP位点上的最小等位基因频率(minor allele frequency, 简记为MAF)在病例组和对照组间有无显著性差异, 进而得出该SNP是否与疾病存在着统计学关联的结论. 伴随着GWAS的快速发展, 常常将GWAS检测出的最显著的SNPs作为CGAS进一步研究的对象, 以缩小目标调查区域.

## 1.3 全基因组关联研究的优势和成效

GWAS具有以下几点优势: (1)它不需要预先获知复杂疾病或性状的生物学通路, 同时它无需家系数据, 可避免家系患病成员DNA测序标本难以获取等限制因素; (2)它有可能发现一些新的通过其他方法识别不出来的基因; (3)它容易形成协作联盟, 除了GWAS外, 他们往往会继续后续分析的合作, 从而有利于发现遗传效应较低的易感基因位点; (4)它可以排除某些特定遗传关联的可能性; (5)它提供两类结构突变的数据—序列偏差及拷贝数变异(copy-number variation, CNV), 可为关联分析提供更稳健可靠的数据.

GWAS取得的成效: 2005年, Science杂志报道了Klein等(2005)对具有年龄相关性的视网膜黄色雀斑进行的GWAS研究, 这也是第一项公开发表的利用商业化基因芯片进行复杂疾病GWAS的结果, 并成功发现一个与视网膜黄色雀斑有显著关联的基因CFH. 随后在全世界范围内, 一系列针对人类复杂疾病或性状的GWAS相继开展, 有关的研究成果陆续见诸报道. 在过去的8年中, GWAS在肿瘤性疾病、内分泌系统疾病、心血管系统疾病、自身免疫性疾病、神经精神类疾病、皮肤病、感染性疾病等诸多复杂疾病领域以及肥胖、皮肤色素、血脂等人类性状方面取得了突飞猛进的研究成果, 为众多复杂疾病和性状下一步的基因诊断及个体化治疗奠定了理论基础. 根据美国国立卫生研究院(National Institutes of Health, NIH)下辖的国家癌症研究所(National Cancer Institute, NCI)以及国家人类基因组研究所(National Human Genome Research Institute, NHGRI)公布在其网站上的GWAS汇总记录(<http://www.genome.gov/gwastudies/>), 截止2013年9月21日, 各国科学家已经对300余种复杂疾病和性状开展1900多项GWAS, 发现11299多个与疾病或性状关联的SNPs, 确定的疾病或性状关联的易感基因或位点超过1000个, 公开报道的GWAS研究成果的数量呈逐年递增态势(图1).

## 1.4 全基因组关联研究的发展现状

2006年以来, 随着人类基因组计划(The Human Genome Project, 简称HGP)和国际人类基因组单体型图计划(The International HapMap Project, 简称HapMap计划)的相继

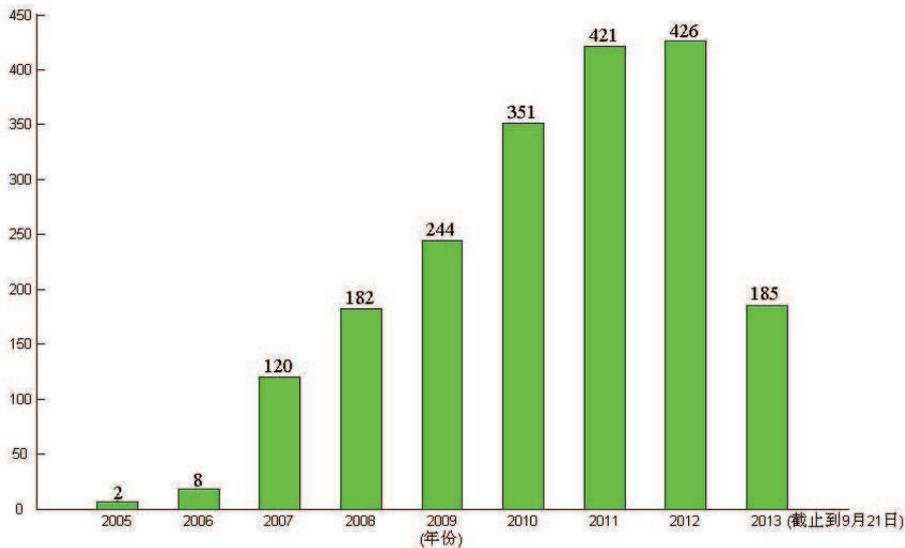


图1 2005年~2013年GWAS研究成果走势图

完成, 以及高通量基因分型技术的飞速发展, 特别是近5年来下一代测序(next-generation sequencing, NGS)技术的出现和应用<sup>[2]</sup>, 使GWAS向着更深更广的领域发展. 当前GWAS的主要工作思路是针对全基因组、外显子组或此前GWAS已识别的易感基因内部进行高通量测序及关联分析, 进而在额外的独立人群中验证, 以期发现复杂疾病或性状的新的低频和罕见变异.

## §2. 全基因组关联研究中存在的问题和挑战

### 2.1 群体分层

如所周知, 关联研究的结果容易受到混杂因素的影响, 而群体分层(population stratification)是导致病例-对照设计下的关联研究中出现假阳性结果的一种最重要的混杂因素. 它是指由于祖先或遗传血统的不同导致病例组与对照组中存在着等位基因频率的系统性差异, 使得许多GWAS的结果难以在不同人群的研究中得到重复或验证. 比如, 在美国的印第安人和高加索人中II型糖尿病的患病率有很大差异. 印第安人的II型糖尿病的患病比率较高, 如果利用印第安人和白人的混合人群进行II型糖尿病的GWAS, 那些在印第安人中频率较高的等位基因型将有可能被认为是与疾病相关联, 而实际上这些等位基因型与II型糖尿病并不一定相关联<sup>[3]</sup>.

### 2.2 多重检验

多重检验(Multiple testing)是GWAS需要面对和处理的一个重要问题. 由于GWAS通常是对基因组上的50万~300万个SNPs同时进行假设检验, 若我们以普通的假设检验中

经常采用的显著性水平 $\alpha = 0.05$ 来判断每一项检验中的SNPs是否与疾病存在着统计学关联, 将会导致第I类错误率增大而带来许多假阳性的关联结果. 尽管统计上有许多控制多重检验的方法, 比如Bonferroni校正法<sup>[4]</sup>、递减的Bonferroni校正法<sup>[5]</sup>、模拟计算校正法(permutation correction)<sup>[6]</sup>、控制假阳性率(false discovery rate, FDR)法<sup>[7]</sup>等, GWAS经常采用的是Bonferroni校正方法, 这是为了尽量降低假阳性率<sup>[8]</sup>.

### §3. 统计分析方法

#### 3.1 遗传模型

定义 $D$ 为一个随机的示性变量,  $D = 1$ 表示个体得病,  $D = 0$ 表示个体不得病. 记 $K = P(D = 1)$ 为整个人群中该疾病的患病率. 假设某SNP位点上的等位基因为 $G$ 和 $g$ , 则该位点上基因型有三种:  $GG$ 、 $Gg$ 和 $gg$ . 不失一般性, 我们假设 $G$ 为风险等位基因, 记 $\theta = P(G)$ 为整个人群中 $G$ 的频率. 将三种基因型的外显率(penetrance)分别记作 $f_0 = P(D = 1|gg)$ ,  $f_1 = P(D = 1|Gg)$ 和 $f_2 = P(D = 1|GG)$ . 假定在整个人群中基因型的分布概率满足哈代-温伯格平衡定律(Hardy-Weinberg equilibrium, HWE), 即 $P(GG) = \theta^2$ ,  $P(Gg) = 2\theta(1 - \theta)$ ,  $P(gg) = (1 - \theta)^2$ . 以基因型 $gg$ 为参照基因型(假设 $f_0 > 0$ ), 基因型 $GG$ 和 $Gg$ 的相对风险(genotype relative risk)为 $\lambda_1 = f_1/f_0$ ,  $\lambda_2 = f_2/f_0$ . 显然有 $K = f_0P(gg) + f_1P(Gg) + f_2P(GG)$ . 记 $p_0 = P(gg|D = 1) = f_0P(gg)/K$ ,  $p_1 = P(Gg|D = 1) = f_1P(Gg)/K$ ,  $p_2 = P(GG|D = 1) = f_2P(GG)/K$ ,  $q_0 = P(gg|D = 0) = (1 - f_0)P(gg)/(1 - K)$ ,  $q_1 = P(Gg|D = 0) = (1 - f_1)P(Gg)/(1 - K)$ ,  $q_2 = P(GG|D = 0) = (1 - f_2)P(GG)/(1 - K)$ . 遗传关联研究就是检验零假设 $H_0 : p_i = q_i, i = 0, 1, 2$ , 即病例组与对照组中基因型的分布无差异. 对立假设可表示为 $H_1 : p_i \neq q_i, i = 0, 1, 2$ . 该假设检验问题等价于

$$H_0 : f_2 = f_1 = f_0 = K \leftrightarrow H_1 : f_2 \geq f_1 \geq f_0, \quad f_2 > f_0.$$

常用的三种遗传模型是: 隐性模型(recessive model)、显性模型(dominant model)和可加模型(additive model), 它们刻画了基因型对表现型(疾病或性状)的影响. 在 $H_1$ 下(该SNP与疾病存在关联), 上面的三种遗传模型可以用基因型的外显率或相对风险来表示: 隐性模型( $f_0 = f_1 < f_2$ 或 $\lambda_1 = 1 < \lambda_2$ )、显性模型( $f_0 < f_1 = f_2$ 或 $1 < \lambda_1 = \lambda_2$ )、可加模型( $f_1 = (f_0 + f_2)/2$ 或 $2\lambda_1 = 1 + \lambda_2$ ). 因此, 对立假设是有序的或受限制的.

#### 3.2 多阶段设计

GWAS常采用多阶段设计(multiple-stage design)的策略来进行基因测序和统计检验, 其中以两阶段设计(two-stage design)最为多见. 第一阶段: 先在部分的病例和对照样本(从全部样本中按照设定比例 $\pi$ 随机抽取, 一般只考虑 $0 < \pi \leq 0.5$ 的情况)中对全基因组范围选择的所有SNPs进行测序, 根据第一阶段统计检验的结果筛选出最显著的一小部分SNPs(几十个到数百个不等); 第二阶段: 在余下的病例和对照样本中对第一阶段选出的SNPs进行测序, 然后结合两个阶段的结果进行联合分析. 类似地可以得出三阶段设计或更高阶段设

计. 相比于一次性对全部样本的全部SNPs进行测序的一阶段设计, 合理构建的两阶段或多阶段设计在保持统计功效的前提下, 不仅可以大大地降低GWAS中基因测序的工作量和费用, 而且能够对两个阶段中分别报告有显著性的SNPs的结果进行比较和验证分析, 从而提高全基因组关联研究的效率. Skol等(2006)指出在两阶段GWAS中将两个阶段的检验统计量加权联合分析(joint analysis)方法比重复分析(replication-based analysis)方法具有更高的统计功效. 注意到, Skol等(2006)给出的联合分析方法是基于等位基因的检验统计量来实施的, 它只有当潜在的遗传模型为可加模型时才有比较高的功效, 但通常情况下我们事先并不知道与疾病关联的SNP位点的遗传模型, 如果假定的遗传模型不正确就可能误报或漏报某些显著的关联结果. Wang等(2006)考虑总的测序费用等因素, 提出了一个最优两阶段设计(optimal two-stage design)的理论框架. Yu等(2007)基于第一阶段的结果, 自适应地来计算第二阶段验证研究所需的样本量. 潘东东等(2011)考虑到遗传模型不确定性, 提出了基于MAX3(隐性、显性、可加模型下Cochran-Armitage趋势性检验统计量的最大值)的两阶段设计及分析方法.

### 3.3 群体分层的纠正方法

纠正GWAS群体分层的方法有很多种, 比如全基因组控制(genomic control, GC)<sup>[13]</sup>, 结构关联(structured association, SA)<sup>[14, 15]</sup>, 主成分分析(PCA)<sup>[16]</sup>, 多尺度分析<sup>[17]</sup>, 偏最小二乘<sup>[18]</sup>, 方差分量模型<sup>[19]</sup>等. 下面对其中的全基因组控制、结构关联和主成分分析法做简要介绍.

#### 3.3.1 全基因组控制

全基因组控制方法假设群体分层将关联分析检验统计量放大了常数因子 $\lambda$ 倍, 通过从所检测的SNPs中选取一部分与复杂疾病没有关联的遗传标记并比较它们在病例组与对照中等位基因的频率差异来估计群体分层效应, 然后再从关联检验统计量中移除这一效应<sup>[13, 20]</sup>. 具体操作是对病例-对照设计下 $2 \times 3$ 的基因型数据表采用Cochran-Armitage趋势性检验<sup>[21]</sup>或在可加遗传模型假设下对 $2 \times 2$ 的等位基因数据表采用 $\chi^2$ 检验<sup>[22]</sup>来计算所选的中性遗传标记的 $\chi^2$ 值, 然后取这些 $\chi^2$ 统计量的中位数除以0.456后的值<sup>[13]</sup>作为群体分层偏离系数或方差膨胀因子 $\lambda$ 的估计(若 $\hat{\lambda} \approx 1$ 表明不存在群体分层; 若 $\hat{\lambda} > 1$ 表明存在群体分层或其他的混杂因素). 后续对全基因组范围内的SNP进行关联检验时, 将各SNP处计算的 $\chi^2$ 统计量除以前面估计出来的方差膨胀因子 $\hat{\lambda}$ 来实现对潜在的群体分层混杂因素的校正.

#### 3.3.2 结构关联

Pritchard等(2000a)提出了SA方法. 它利用多个遗传标记处的基因型数据, 首先采用基于模型的贝叶斯聚类算法来推断群体结构并指定各个样本个体所归属的子群体, 然后在各子群体中分别进行关联检验, 最后将各子群体中关联检验的结果进行综合来得到总体的关联研究结果. 具体如下: 假设一个具有 $W$  ( $W$ 可能是未知的)个类(子群体)的

模型, 每个子群体由所选遗传标记处的一系列等位基因频率来刻画并区分. 令  $X$  表示每个样本个体的基因型数据,  $Y = (y_1, y_2, \dots, y_W)$  表示每个个体的祖先群体的可能性 ( $y_i$  为该个体的基因组源自第  $i$  个子群体的比例),  $Z$  表示各子群体中选取的标记位点处(未知的)等位基因频率. 假定各群体中HWE成立且各遗传标记处于连锁平衡, 由此完全确定了条件概率分布  $P(X|Y, Z)$ . 基于观测的基因型数据  $X$  及先验信息  $P(Y)$  (假设每个个体都来自单个群体时,  $y_i = 1/W, i = 1, \dots, W$ ) 和  $P(Z)$  (取为Dirichlet分布), 可得到后验分布  $P(Y, Z|X) \propto P(X|Y, Z)P(Y)P(Z)$  以及  $P(W|X) \propto P(X|W)P(W)$ . 再利用Markov chain Monte Carlo (MCMC) 方法和Gibbs抽样方法产生的后验分布的随机样本来对  $W, Y$  和  $Z$  进行推断, 即可同时估计出子群体的数目、每个子群体中所考察的遗传标记处等位基因的频率、每个样本个体的遗传背景(按概率大小归类到相应的子群体中)<sup>[15]</sup>. 该算法的软件Structure可从Pritchard实验室的网站上自由下载(<http://pritchardlab.stanford.edu/structure.html>), 由于SA方法计算量大、成本高, 使得它在GWAS中的应用受到一定限制.

### 3.3.3 主成分分析

Price等(2006)提出了基于PCA的EIGENSTRAT方法, 其主要分为三个步骤: (i) 对基因型数据应用PCA得到个体的“祖先位置”. 假设我们有  $N$  个个体, 每个个体有  $M$  个SNPs. 令  $g_{ij}$  为第  $j$  个个体第  $i$  个SNP位点上的基因型并以此构造  $M \times N$  的基因型数据矩阵. 将该矩阵中的元素按照行进行标准化(每个元素都减去它所在行的均值后除以  $\sqrt{p_i(1-p_i)}$ , 其中  $p_i = (1 + \sum_j g_{ij}) / (2 + 2N)$ ), 记变换后的矩阵为  $X$  并计算其  $N \times N$  的协方差矩阵  $\Psi$  ( $\Psi_{jj'}$  表示  $X$  的第  $j$  列与第  $j'$  列之间的协方差). 定义  $\Psi$  的第  $k$  个特征向量为遗传变异的第  $k$  个轴, 则第  $j$  个个体沿着遗传变异的第  $k$  个轴方向的世系  $a_{jk}$  等于第  $k$  个特征向量的第  $j$  个元素; (ii) 选择前10个主成分并计算线性回归的残差, 即沿每一条坐标轴对基因型和表现型中可归咎于世系的部分进行连续性地调整. 令  $a_j$  为第  $j$  个个体沿遗传变异某一给定的轴上的世系, 定义  $g_{ij}^* = g_{ij} - \gamma_i a_j$ , 其中  $\gamma_i = \sum_j a_j g_{ij} / \sum_j a_j^2$  是以世系来对在 第  $i$  个SNP处具有有效基因型的个体做预测时的回归系数. 类似地, 对遗传变异的每一条轴和样本个体的表现型也进行调整; (iii) 对经过世系调整后的基因型和表现型数据计算关联检验的  $\chi^2$  统计量. 由于EIGENSTRAT方法具有计算简便、耗时短等多方面的优势, 可以同时处理数以万计的SNPs位点, 逐渐成为GWAS中应用最为广泛的群体分层校正方法. 注意Price等提出的主成分就是特征向量, 使很多实际工作者较难理解, Li和Yu (2008)应用多尺度方法 (multidimensional scaling, MDS)<sup>[23]</sup> 证明了该特征向量也是主成分, 并提出基于距离回归的思想构建一个伪  $F$  统计量来对测序样本进行遗传背景比较, 并通过选择全部主成分的一个最优子集来实现对群体分层的评估及校正<sup>[24]</sup>.

### 3.4 单位点检验

假设在一项GWAS中征集了  $r$  个病例个体和  $s$  个对照个体并对他们的基因型进行测序. 总样本量为  $n = r + s$ . 记病例组和对照组中具有基因型 (gg, Gg, GG) 的人数分别为  $(r_0, r_1,$

$r_2$ )和 $(s_0, s_1, s_2)$ . 将基因型数据汇总如表1所示.

表1 病例组和对照组中基因型的频数

	gg	Gg	GG	和
病例	$r_0$	$r_1$	$r_2$	$r$
对照	$s_0$	$s_1$	$s_2$	$s$
和	$n_0$	$n_1$	$n_2$	$n$

### 3.4.1 等位基因频率差检验

记患病群体和健康群体中风险等位基因的频率分别为 $\theta_1$ 和 $\theta_2$ . 基于等位基因的频率差检验统计量为

$$Z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\left(\frac{1}{2r} + \frac{1}{2s}\right)[\hat{\theta}(1 - \hat{\theta})]},$$

其中 $\hat{\theta}_1 = (2r_2 + r_1)/2r$ ,  $\hat{\theta}_2 = (2s_2 + s_1)/2s$ ,  $\hat{\theta} = (2n_2 + n_1)/2n$ . 当 $\min\{r, s\} \rightarrow \infty$ 且 $r/n \rightarrow \xi$  ( $0 < \xi < \infty$ ),  $Z$ 在 $H_0$ 下渐近服从 $N(0, 1)$ .

### 3.4.2 Cochran-Armitage趋势性检验

考虑到剂量效应(Dose effect), 即风险等位基因个数的增加会增加一个人得病的风险, Cochran-Armitage趋势性检验统计量<sup>[21, 22]</sup>为

$$\text{CATT}(x) = \frac{\left[ \sum_{i=0}^2 x_i \left( \frac{r_i}{r} - \frac{s_i}{s} \right) \right]}{\sqrt{\frac{n}{rs} \left[ \sum_{i=0}^2 x_i^2 \frac{n_i}{n} - \left( \sum_{i=0}^2 x_i \frac{n_i}{n} \right)^2 \right]}},$$

其中 $(x_0, x_1, x_2) = (0, x, 1)$ 分别代表基因型(gg, Gg, GG)的得分.  $x = 0, 1$ 分别为隐性模型和显性模型的最优得分, 而 $x = 0.5$ 为可加模型下的最优得分(即若 $x$ 正好对应于所考察的SNP的真实遗传模型, 则CATT( $x$ )是检验该SNP与疾病存在关联性的功效最高的检验<sup>[25]</sup>). 当 $\min\{r, s\} \rightarrow \infty$ 且 $r/n \rightarrow \xi$  ( $0 < \xi < \infty$ ), CATT( $x$ )在 $H_0$ 下渐近服从 $N(0, 1)$ .

### 3.5 稳健的单位点检验

GWAS中常用的检验统计量, 比如可加遗传模型下的趋势性检验统计量, 其功效都依赖于疾病位点处潜在的遗传模型. 然而, 在实际中我们只知道一族遗传学家们认可的遗传模型但不清楚其中哪一个是真实的, 因此可以得到对应于遗传模型的一族检验统计量. 如果遗传模型指定有误而使用族中其它的检验统计量, 将会使假阴性和假阳性的结果显著增大. 针对遗传模型不确定性问题, 稳健检验统计量则是一个比较好的选择<sup>[26]</sup>.

### 3.5.1 皮尔逊 $\chi^2$ 检验

三项分布:  $(r_0, r_1, r_2) \sim \text{Mul}(r, p_0, p_1, p_2)$ ,  $(s_0, s_1, s_2) \sim \text{Mul}(s, q_0, q_1, q_2)$ . 皮尔逊(Pearson)  $\chi^2$ 统计量为

$$\text{CHI2} = \sum_{i=0}^2 \frac{(r_i - n_i r/n)^2}{n_i r/n} + \sum_{i=0}^2 \frac{(s_i - n_i s/n)^2}{n_i s/n},$$

在零假设 $H_0$ 成立时, 该CHI2检验统计量渐近服从自由度为2的中心卡方分布.

### 3.5.2 限制似然比检验

限制似然比检验(constrained likelihood ratio test, CLRT)由Wang和Sheffield (2005)提出, 其检验统计量的形式如下

$$\text{CLRT} = 2 \left[ \max_{\{f_0 \leq f_1 \leq f_2 \text{ 且 } f_0 < f_2\} \text{ 或 } \{f_0 \geq f_1 \geq f_2 \text{ 且 } f_0 > f_2\}} l_2(f_0, f_1, f_2) - l_0 \right],$$

其中 $l_2(f_0, f_1, f_2) = \sum_{i=0}^2 r_i \log f_i + s_i \log(1 - f_i)$ ,  $l_0 = r \log \hat{f}_0 + s \log(1 - \hat{f}_0)$ ,  $\hat{f}_0 = r/n$ .

CLRT的具体构造方式为: (i) 当外显率的自然估计 $\hat{f}_i = r_i/n_i$ ,  $i = 0, 1, 2$ 有递增或递减的顺序, 则直接将它们的估计值代入 $l_2$ ; (ii) 当 $(r_0/n_0, r_1/n_1, r_2/n_2)$ 没有按顺序排列时, 先在隐性模型下计算 $\hat{f}_0 = \hat{f}_1 = (r_0 + r_1)/(n_0 + n_1)$ 和 $\hat{f}_2 = r_2/n_2$ 再把它们代入 $l_2$ 中得到隐性模型下的CLRT检验统计量(记为CLRTR), 然后在显性模型下计算 $\hat{f}_0 = r_0/n_0$ 和 $\hat{f}_1 = \hat{f}_2 = (r_1 + r_2)/(n_1 + n_2)$ 再把它们代入CLRT表达式得到显性模型下的检验统计量(记为CLRTD), 最后计算限制似然比检验统计量为 $\text{CLRT} = \max(\text{CLRTR}, \text{CLRTD})$ . 在零假设下, CLRT的渐近分布是多个卡方分布的一个混合形式.

### 3.5.3 MAX

MAX检验统计量的形式为 $\text{MAX} = \max_{x \in [0,1]} |\text{CATT}(x)|$ , 即取有效得分域内CATT统计量的最大值. 要获得MAX在 $H_0$ 下的渐近分布是比较困难的, 但可以采用Bootstrap方法来模拟其近似分布<sup>[28]</sup>, 李启寨等(2010)给出了另一种有效的计算公式来近似MAX的P-值.

### 3.5.4 最大最小效率稳健检验

原始的最大最小效率稳健检验(maximin efficiency robust test, MERT)由Gastwirth在1985年提出<sup>[30]</sup>. 当多个对立假设模型都看似合理时可以用它来得到比较稳健的的检验统计量. Zheng等(2006)把它应用来检验SNP与复杂疾病之间的关联, 检验统计量如下

$$\text{MERT} = \frac{\text{CATT}(0) + \text{CATT}(1)}{\sqrt{2(1 + \rho_{0,1})}},$$

其中  $\rho_{0,1} = p_0 p_2 / [\sqrt{p_0(1-p_0)}\sqrt{p_2(1-p_2)}]$  为  $H_0$  下 CATT(0) 和 CATT(1) 之间的相关系数 (在实际计算中,  $p_i$  用  $\hat{p}_i = n_i/n$ ,  $i = 0, 1, 2$  来进行估计). 当  $\min\{r, s\} \rightarrow \infty$  且  $r/n \rightarrow \xi$  ( $0 < \xi < \infty$ ) 时, MERT 在  $H_0$  下渐近服从  $N(0, 1)$ .

### 3.5.5 MAX3

Freidlin 等(2002)提出了  $\text{MAX3} = \max\{|\text{CATT}(0)|, |\text{CATT}(1/2)|, |\text{CATT}(1)|\}$ , 即隐性、显性和可加遗传模型下三个 CATT 统计量的最大值. MAX3 考虑了常用的三种遗传模型, 因此它比任意单个 CATT 具有更好的稳健性. 这三个 CATT 统计量在  $H_0$  下的相关系数为 ( $\rho_{0,1}$  在前面已经给出)

$$\rho_{0,1/2} = \frac{p_2(p_1 + 2p_0)}{\sqrt{p_2(1-p_2)}\sqrt{(p_1 + 2p_2)p_0 + (p_1 + 2p_0)p_2}},$$

$$\rho_{1/2,1} = \frac{p_0(p_1 + 2p_2)}{\sqrt{p_0(1-p_0)}\sqrt{(p_1 + 2p_2)p_0 + (p_1 + 2p_0)p_2}},$$

再根据 CATT(0), CATT(1/2), CATT(1) 三者之间的线性关系  $\text{CATT}(1/2) = \omega_0 \text{CATT}(0) + \omega_1 \text{CATT}(1)$ , 其中  $\omega_0 = (\rho_{0,1/2} - \rho_{0,1}\rho_{1/2,1})/(1 - \rho_{0,1}^2)$ ,  $\omega_1 = (\rho_{1/2,1} - \rho_{0,1}\rho_{0,1/2})/(1 - \rho_{0,1}^2)$ , 以及 (CATT(0), CATT(1)) 在  $H_0$  下渐近服从的二元正态分布  $f(z_0, z_1; \Sigma)$  (其均值向量为  $\mathbf{0}$ , 协方差阵  $\Sigma$  的主对角元素为 1 非对角元素为  $\rho_{0,1}$ ), 即可实现 MAX3 的  $P$ -值计算  $P(\text{MAX3} > t) = 1 - P(\text{MAX3} < t)$ , 而  $P(\text{MAX3} < t)$  可简化为如下的二重积分来计算

$$\begin{aligned} & P_{H_0}(\text{MAX3} < t) \\ &= P(|\text{CATT}(0)| < t, |\text{CATT}(1/2)| < t, |\text{CATT}(1)| < t) \\ &= P(|\text{CATT}(0)| < t, |\omega_0 \text{CATT}(0) + \omega_1 \text{CATT}(1)| < t, |\text{CATT}(1)| < t) \\ &= 2 \int_0^{t(1-\omega_1)/\omega_0} \int_{-t}^t f(z_0, z_1; \Sigma) dz_1 dz_0 + 2 \int_{t(1-\omega_1)/\omega_0}^t \int_{-t}^{(t-\omega_0 z_0)/\omega_1} f(z_0, z_1; \Sigma) dz_1 dz_0 \\ &= 2 \left[ \int_0^{t(1-\omega_1)/\omega_0} \Phi\left(\frac{t - \rho_{0,1} z_0}{\sqrt{1 - \rho_{0,1}^2}}\right) \phi(z_0) dz_0 \right. \\ &\quad \left. + \int_{t(1-\omega_1)/\omega_0}^t \Phi\left(\frac{(t - \omega_0 z_0)/\omega_2 - \rho_{0,1} z_0}{\sqrt{1 - \rho_{0,1}^2}}\right) \phi(z_0) dz_0 - \int_0^t \Phi\left(\frac{-t - \rho_{0,1} z_0}{\sqrt{1 - \rho_{0,1}^2}}\right) \phi(z_0) dz_0 \right], \end{aligned}$$

其中  $\phi(\cdot)$  和  $\Phi(\cdot)$  分别为标准正态分布的密度函数和累积分布函数. 李启寨等(2008)给出了一种有效的计算公式近似 MAX3 的  $P$ -值, 同时, 他们还给出了在有协变量时如何计算 MAX3 的得分检验统计量.

### 3.5.6 MIN2

WTCCC (2007) 提出并应用 MIN2 来检验七种疾病与 SNPs 之间的关联. MIN2 是取 Pearson 卡方检验 (记作  $T_1$ ) 的  $P$ -值 (记作  $P_1$ ) 与可加模型下趋势性检验的  $P$ -值 (记作  $P_2$ ) 这两

者中的较小值来对SNP进行检验并以此作为最终评价SNPs与疾病关联性的依据, 即

$$\text{MIN2} = \min\{P_1, P_2\}.$$

记 $T_2 = [\text{CATT}(1/2)]^2$ , 并令 $F_i$ 为 $T_i, i = 1, 2$ 的累积分布函数, MIN2在 $H_0$ 下的分布函数 $F$ 为

$$\begin{aligned} F(c) &= \mathbf{P}_{H_0}(\text{MIN2} < c) = 1 - \mathbf{P}_{H_0}(P_1 > c, P_2 > c) \\ &= 1 - \mathbf{P}_{H_0}(T_1 < F_1^{-1}(1-c), T_2 < F_2^{-1}(1-c)). \end{aligned}$$

Joo等(2009)进一步给出了计算MIN2检验的渐近 $P$ -值的显表达式如下

$$\begin{aligned} P_{\text{MIN2}} &= F(\text{MIN2}) = 1 - \mathbf{P}_{H_0}(T_1 < F_1^{-1}(1 - \text{MIN2}), T_2 < F_2^{-1}(1 - \text{MIN2})) \\ &= \frac{1}{2}e^{-F_2^{-1}(1-\text{MIN2})/2} + \frac{1}{2}\text{MIN2} \\ &\quad - \frac{1}{2\pi} \int_{F_2^{-1}(1-\text{MIN2})}^{-2\log(\text{MIN2})} e^{-v/2} \arcsin\left(\frac{2F_2^{-1}(1 - \text{MIN2})}{v} - 1\right) dv. \end{aligned}$$

### 3.5.7 遗传模型选择检验

Zheng等(2008)提出了一种遗传模型选择(genetic model selection, GMS)的稳健检验方法. 它分为两个步骤: 第一步使用Song和Elston(2006)提出的哈代-温伯格不平衡趋势性检验(Hardy-Weinberg disequilibrium trend test, HWDTT)来测出SNP潜在的遗传模型, 这里HWDTT统计量的形式如下

$$\text{HWDTT} = \frac{\left[\hat{p}_2 - \left(\hat{p}_2 + \frac{1}{2}\hat{p}_1\right)^2\right] - \left[\hat{q}_2 - \left(\hat{q}_2 + \frac{1}{2}\hat{q}_1\right)^2\right]}{\sqrt{\left(\frac{1}{r} + \frac{1}{s}\right)\left(\frac{1}{16n^4}\right)[2n - (2n_2 + n_1)]^2(2n_2 + n_1)^2}},$$

其中 $\hat{p}_i = r_i/r, \hat{q}_i = s_i/s, i = 1, 2$ . 在 $H_0$ 下 $\text{HWDTT} \sim N(0, 1)$ , 取 $c = \Phi^{-1}(0.95) = 1.645$ , 当 $\text{HWDTT} > c$ 时选择隐性模型, 当 $\text{HWDTT} < -c$ 时选择显性模型, 当 $|\text{HWDTT}| < c$ 时则为可加模型. 第二步使用与选择的遗传模型相对应的最优的CATT来进行关联检验. 最后的GMS统计量即为

$$\begin{aligned} \text{GMS} &= \text{CATT}(0)I\{\text{CATT}(0.5) > 0\}I\{\text{HWDTT} > c\} \\ &\quad + \text{CATT}(0.5)I\{\text{CATT}(0.5) > 0\}I\{|\text{HWDTT}| \leq c\} \\ &\quad + \text{CATT}(1)I\{\text{CATT}(0.5) > 0\}I\{\text{HWDTT} < -c\} \\ &\quad - \text{CATT}(1)I\{\text{CATT}(0.5) \leq 0\}I\{\text{HWDTT} > c\} \\ &\quad - \text{CATT}(0.5)I\{\text{CATT}(0.5) \leq 0\}I\{|\text{HWDTT}| \leq c\} \\ &\quad - \text{CATT}(0)I\{\text{CATT}(0.5) \leq 0\}I\{\text{HWDTT} < -c\}, \end{aligned}$$

其中 $I\{\cdot\}$ 为示性函数. Joo等(2010)进一步研究了GMS在 $H_0$ 下的渐近分布. 与MAX3类似, HWDTT渐近地也可表示为CATT(0)和CATT(0.5)的线性组合 $\text{HWDTT} = \mu_0\text{CATT}(0) +$

$\mu_1 \text{CATT}(1)$ , 其中  $\mu_0 = (\rho_0 - \rho_{0,1}\rho_1)/(1 - \rho_{0,1}^2)$ ,  $\mu_1 = (\rho_1 - \rho_{0,1}\rho_0)/(1 - \rho_{0,1}^2)$ , 而  $\rho_0 = \sqrt{(1-\theta)/(1+\theta)}$  为  $\text{CATT}(0)$  与  $\text{HWDTT}$  的渐近零相关系数,  $\rho_1 = -\sqrt{\theta/(2-\theta)}$  为  $\text{CATT}(1)$  与  $\text{HWDTT}$  在零假设成立时的相关系数. 记  $(\text{CATT}(0), \text{CATT}(0.5), \text{HWDTT})$  和  $(\text{CATT}(1), \text{CATT}(0.5), -\text{HWDTT})$  的联合密度函数分别为  $f_1(\mathbf{z}; \Sigma_1)$  和  $f_2(\mathbf{z}; \Sigma_2)$  (均值向量都为  $\mathbf{0}$ , 协方差阵为  $\Sigma_i, i = 1, 2$  的三元正态分布), 根据实际数据计算  $\text{GMS}$  值并取绝对值为  $t, t > 0$ , 可以按如下的公式来计算  $\text{GMS}$  的  $P$ -值

$$\begin{aligned} P_{\text{GMS}} &= 2P_{H_0}(\text{GMS} > t) \\ &= 4[\text{P}(\text{CATT}(0) > t, \text{CATT}(0.5) > 0, \text{HWDTT} > c) \\ &\quad + \text{P}(\text{CATT}(1) > t, \text{CATT}(0.5) > 0, -\text{HWDTT} > c) \\ &\quad + 0.9\text{P}(\text{CATT}(0.5) > t)] \\ &= 3.6 - 3.6\Phi(t) + 4 \int_t^{+\infty} \int_0^{+\infty} \int_c^{+\infty} [f_1(\mathbf{z}; \Sigma_1) + f_2(\mathbf{z}; \Sigma_2)] d\mathbf{z}. \end{aligned}$$

### 3.6 非参数检验

近年来研究人员也提出若干用于  $\text{GWAS}$  的非参数检验方法. 当被考察的性状是二值型时, Schaid 等(2005)提出了一种基于  $U$ -统计量的关联检验方法. 首先分别对病例组和对照组中所有的样本个体的配对采用累加型核函数来计算指定的位点集上基因型的得分, 然后将病例组内与对照组内得到的基因型得分向量进行加权平均后作差, 得到的全局检验统计量仅有 1 个自由度, 因而具有显著的功效优势.

当被考察的性状是连续时, Zhang 等(2010)提出了一种适合单个位点关联分析的广义 Kendall  $\tau$  秩检验方法. 朱文圣等(2012)通过对样本对赋权的方法研究了带协变量调整的关联分析的广义 Kendall  $\tau$  秩检验方法. 李启寨等(2013)从遗传模型的角度提出了一种适合单个位点关联分析的非参数趋势性检验方法, 所给出的非参数趋势性检验方法非常直观, 而且给出了其在原假设下的方差表达式, 以此研究这种非参数检验方法的功效, 并提出了一类对所有潜在遗传模型(可加遗传模型、显性遗传模型、隐性遗传模型)都具有稳健性的非参数检验方法.

### 3.7 贝叶斯因子

贝叶斯因子(Bayes factor, BF)由 Jeffreys 在 1961 年提出<sup>[42]</sup>, 并被应用于  $\text{GWAS}$  中<sup>[33]</sup>. 记观测数据  $\Upsilon$  在  $H_0$  和  $H_1$  下的概率分别记为  $\text{P}(\Upsilon|H_0)$  和  $\text{P}(\Upsilon|H_1)$ , 贝叶斯因子的定义为

$$\text{BF}_{10} = \frac{\text{P}(\Upsilon|H_1)}{\text{P}(\Upsilon|H_0)}.$$

BF 在形式上类似于频率统计学派中的似然比, 但 BF 可以比较两个不同模型, 而似然比一般是比较同一个模型下两个不同的参数值<sup>[43]</sup>. BF 是观测数据对  $H_0$  或  $H_1$  偏好程度的一个度量, 越大的 BF 值越支持数据来自于  $H_1$ , 当 BF 的值为 1 时说明数据  $Y$  在  $H_0$  与  $H_1$  下具

有相等的可能性. Kass和Raftery (1995)仿照似然比检验统计量将BF取自然对数后再乘以2即 $2\ln(\text{BF}_{10})$ , 给出了实际中应用BF的一个评判准则: 当 $2\ln(\text{BF}_{10})$ 在0到2之间, 较弱的证据拒绝 $H_0$ ; 2到6之间, 明确的证据拒绝 $H_0$ ; 6到10之间, 强烈的证据拒绝 $H_0$ ; 大于10时, 非常强烈的证据拒绝 $H_0$ . 假定观测数据 $\Upsilon$ 在 $H_0$ 和 $H_1$ 下的概率密度函数分别是 $p(\Upsilon|\zeta_0, H_0)$ 和 $p(\Upsilon|\zeta_1, H_1)$  (其中 $\zeta_i$ 是 $H_i, i = 0, 1$ 下的参数), 则有

$$\text{BF}_{10} = \frac{P(\Upsilon|H_1)}{P(\Upsilon|H_0)} = \frac{\int p(\Upsilon|\zeta_1, H_1)p(\zeta_1|H_1)d\zeta_1}{\int p(\Upsilon|\zeta_0, H_0)p(\zeta_0|H_0)d\zeta_0},$$

其中 $p(\zeta_i|H_i)$ 是 $H_i, i = 0, 1$ 下的先验密度函数. 注意到, 在高维情形下通常难以得到BF表达式中两个积分项的精确解析结果, 而两种常用的近似方法(拉普拉斯展开和蒙特卡洛抽样)的计算量都非常大, 不便于在GWAS中快速地分析上百万个SNPs. Wakefield (2007)提出了一种近似计算BF的简捷方法, 称作渐近贝叶斯因子(asymptotic Bayes factor, ABF). 它基于如下的logistic回归模型

$$\text{logit}[P(D = 1|\zeta, g_i)] = \beta_0 + \beta_1 g_i, \quad i = 0, 1, 2,$$

其中 $\zeta = (\beta_0, \beta_1)$ ,  $\beta_0$ 为讨厌参数,  $\beta_1$ 为SNP遗传效应的系数,  $g_i, i = 0, 1, 2$ 是依赖于遗传模型、对应基因型(gg, Gg, GG)取值的变量. 这时零假设转化为 $H_0 : \beta_1 = 0$ .  $(\beta_0, \beta_1)$ 的MLE $(\hat{\beta}_0, \hat{\beta}_1)$ 具有渐近分布

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim N_2 \left( \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \begin{bmatrix} i_{00} & i_{01} \\ i_{10} & i_{11} \end{bmatrix}^{-1} \right),$$

其中 $i_{00}, i_{01} = i_{10}, i_{11}$ 为观测的Fisher信息阵中的元素在 $(\hat{\beta}_0, \hat{\beta}_1)$ 处的取值. 作变换 $\beta_0^* = \beta_0 + (i_{01}/i_{00})\beta_1$ , 其估计为 $\hat{\beta}_0^* = \hat{\beta}_0 + (i_{01}/i_{00})\hat{\beta}_1$ , 则有

$$\begin{bmatrix} \hat{\beta}_0^* \\ \hat{\beta}_1 \end{bmatrix} \sim N_2 \left( \begin{bmatrix} \beta_0^* \\ \beta_1 \end{bmatrix}, \begin{bmatrix} i_{00}^{-1} & 0 \\ 0 & (i_{11} - i_{01}^2/i_{00})^{-1} \end{bmatrix}^{-1} \right).$$

将BF中的观测数据 $\Upsilon$ 以估计量 $\hat{\beta}_0^*, \hat{\beta}_1$ 代替, 根据 $\hat{\beta}_0^*$ 与 $\hat{\beta}_1$ 的渐近独立性质, 并假定 $\beta_0^*$ 与 $\beta_1$ 在 $H_1$ 下的先验相互独立, 即 $p(\beta_0^*, \beta_1|H_1) = p(\beta_0^*|H_1)p(\beta_1|H_1)$ , 而 $\beta_0^*$ 在 $H_0$ 和 $H_1$ 下取相同的先验, 即 $p(\beta_0^*|H_0) = p(\beta_0^*|H_1)$ , 从而得到渐近贝叶斯因子

$$\text{ABF}_{10} = \frac{P(\hat{\beta}_0^*, \hat{\beta}_1|H_1)}{P(\hat{\beta}_0^*, \hat{\beta}_1|H_0)} = \frac{\int p(\hat{\beta}_1|\beta_1, H_1)p(\beta_1|H_1)d\beta_1}{p(\hat{\beta}_1|H_0)},$$

这里 $p(\hat{\beta}_1|H_0)$ 是 $N(0, \sigma_0^2)$ 的密度函数,  $p(\hat{\beta}_1|\beta_1, H_1)$ 是 $N(\beta_1, \sigma_0^2)$ 的密度函数, 其中 $\sigma_0^2 = (i_{11} - i_{01}^2/i_{00})^{-1}$ . 若取 $\beta_1$ 在 $H_1$ 下的先验为 $N(0, \sigma_1^2)$ , 进一步有

$$\text{ABF}_{10} = \sqrt{\frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2}} \exp \left( \frac{\hat{\beta}_1^2}{2\sigma_0^2} \times \frac{\sigma_1^2}{\sigma_0^2 + \sigma_1^2} \right),$$

其中 $\hat{\beta}_1^2/\sigma_0^2$ 就是检验 $H_0: \beta_1 = 0$ 的Wald统计量. 计算 $ABF_{10}$ 依赖于 $\beta_1$ 的先验分布 $N(0, \sigma_1^2)$ 的选取, Wakefield (2009)给出了一种选择 $\sigma_1^2$ 的方法: 假定优势比 $\exp(\beta_1)$ 以一个很小的概率 $p_{\beta_1}$ 大于指定的上界 $U_{\beta_1}$ , 即 $P(\beta_1 \leq \ln(U_{\beta_1})) = 1 - p_{\beta_1}$ , 解得 $\sigma_1 = \ln(U_{\beta_1})/\Phi^{-1}(1 - p_{\beta_1})$ . 当取 $p_{\beta_1} = 0.05$ ,  $U_{\beta_1} = 1.5$ 时有 $\sigma_1^2 \approx 0.061$ .

## §4. 典型案例分析

### 4.1 Wellcome Trust Case Control Consortium

2005年, 英国50个研究机构合作成立了一个研究协会, The Wellcome Trust Case Control Consortium (WTCCC). 该协会旨在寻找与复杂疾病相关联的突变位点. 迄今为止, 该协会已经发现近90个与II型糖尿病, 冠心病和类风湿性关节炎等人类复杂疾病相关联的SNPs. 2007年, WTCCC通过对7种常见复杂疾病(双向性情感障碍、冠心病、克罗恩病、高血压、类风湿性关节炎、I型糖尿病、II型糖尿病)共计14000个病例样本(每种疾病都有2000个病例样本)和3000个共享对照样本进行大规模的GWAS(17000个样本全部为生活在大不列颠岛的欧洲白种人, 总共测序了500568个SNP). 在经过Bonferoni校正的显著性水平( $5 \times 10^{-7}$ )下确定了6种疾病的24个独立的致病基因位点(双向情感障碍和冠心病各1个、克罗恩病9个、类风湿性关节炎3个、I型糖尿病7个、II型糖尿病3个), 另外发现了58个额外的易感位( $P$ -值介于 $1 \times 10^{-5}$ 至 $5 \times 10^{-7}$ 之间)<sup>[33]</sup>, 这些疾病关联位点多数被其它研究团队独立进行的GWAS所证实<sup>[47-49]</sup>.

### 4.2 牛皮癣和精神分裂症

GWAS在中国起步较欧美国家晚, 但经过国内科学家的不懈努力, 近年来也取得了一些研究成果. 下面以牛皮癣和精神分裂症的GWAS为例. 2009年, 张学军团队完成了中国人牛皮癣(psoriasis, 学名为银屑病)的GWAS<sup>[50]</sup>. 他们对6860位银屑病患者和8472位正常对照个体进行了基因测序, 统计检验的结果发现位于染色体1q21上的编码晚期角质化包膜(late cornified envelope, LCE)基因簇与牛皮癣之间有显著的遗传关联性, 提示表皮角质形成细胞的终末分化在牛皮癣的发病机制中有重要作用, 该研究成果已在欧洲人种的研究中得到了验证<sup>[51]</sup>. 此外, 该研究还证实了两个以往报道的牛皮癣易感基因MHC和IL2B. 2010年该团队在前期样本数据的基础上进一步扩大了验证样本量(总共测序了11605个患者和17107个对照及254个核心家系), 并与多家国际牛皮癣研究机构合作, 在欧美人群中对研究结果进行验证分析, 新发现了6个同牛皮癣相关联的易感基因位点<sup>[52]</sup>.

精神分裂症(schizophrenia)是一种严重的精神疾病. 研究者们认为遗传、幼年环境、神经科学及心理与社会历程是导致精神分裂症的重要因素. 2010年, 李涛等(2010)为了验证对高加索人群的GWAS所发现的与精神分裂症有关联的7个SNP<sup>[54]</sup>在中国汉族人群中是否也有显著关联, 采集了2496个精神分裂症患者和5184个正常对照的基因型数据, 结果发现有4个SNPs在汉族人群中不是多态的, 另外三个SNPs在汉族群体中与精神分裂症有显著

的关联性. 2011年, 岳伟华等(2011)在汉族人群中对精神分裂症进行GWAS, 发现基因组区域6p21-p22.1和11p11.2区域是精神分裂症的易感区域.

## §5. GWAS展望

GWAS在人类复杂疾病和性状的机理研究方面已经取得了许多引人瞩目的成就, 许多与疾病/性状相关联的易感基因位点陆续被发现. 但我们也应当清醒地认识到, 对复杂疾病和性状的GWAS仅仅是揭示复杂疾病/性状潜在的生物学机制漫长道路上的一个开端<sup>[56, 57]</sup>, 与后续的个性化诊断及治疗还有相当长的一段距离. 注意到, 以往的GWAS主要基于“常见疾病/常见变异”假设来进行, 这仅适用于与疾病/性状相关联的常见变异研究. 越来越多的研究结果表明低频多态性或稀有变异, 比如拷贝数多态性(CNV), 对于常见变异无法解释的遗传可能性起着很大一部分的作用<sup>[58, 59]</sup>, 特别是随着下一代测序技术的快速发展使得低成本获取稀有变异的海量数据逐渐成为现实, 在“常见疾病/稀有变异”(common disease/rare variant, CDRV)假设下发展新的高效统计检验方法来分析此类数据变得迫切而又富有挑战性<sup>[60-62]</sup>. 最近研究人员提出基于家系数据的GWAS并采用经典的传递不平衡检验(transmission disequilibrium test, TDT)来进行统计分析<sup>[63-65]</sup>, 另一种策略是使用基于家系的关联分析对前期GWAS发现的易感位点进行验证分析<sup>[66]</sup>, 这一方法已成功地应用到前列腺癌等人类严重疾病的致病机理研究中<sup>[67]</sup>. 同时, 也有不少学者提出将临床试验研究中的荟萃分析(meta-analysis)的思想和方法用于GWAS, 即组合针对同一种疾病或性状进行的多项GWAS的数据或结果来提升关联检验的功效, 这就有可能找出那些单个GWAS未能发现但确实对疾病的风险具有细微效应的遗传突变<sup>[68]</sup>, 而如何有效地剔除各个GWAS所用的样本中潜在的混杂因素也是一个不可回避的问题<sup>[69]</sup>. 另外, 考虑SNP之间交互效应的多位点检验方法也逐步兴起, 但其中涉及到超高维巨型计算、多重检验等问题还有待研究人员给出更好的解决办法<sup>[70, 71]</sup>. 总之, GWAS在世界范围内仍处于不断发展中, 它需要整合多个学科的专长以及临床医师、流行病学家、遗传学家、统计学家、计算机科学家等多方面的力量来共同协作, 给科研工作者们提供了新的机遇和挑战.

## 参 考 文 献

- [1] Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C. and Hoh, J., Complement factor H polymorphism in age-related macular degeneration, *Science*, **308(5720)**(2005), 385-389.
- [2] Luo, L., Boerwinkle, E. and Xiong, M., Association studies for next-generation sequencing, *Genome Research*, **21(7)**(2011), 1099-1108.
- [3] Knowler, W.C., Williams, R.C., Pettitt, D.J. and Steinberg, A.G.,  $Gm^{3;5,13,14}$  and type 2 diabetes mellitus: an association in American Indians with genetic admixture, *American Journal of Human Genetics*, **43(4)**(1988), 520-526.

- [4] Bland, J.M. and Altman, D.G., Multiple significance tests: the Bonferroni method, *British Medical Journal*, **310(6973)**(1995), 170.
- [5] Holland, B.S. and Copenhaver, M.D., An improved sequentially rejective Bonferroni test procedure, *Biometrics*, **43(2)**(1987), 417–423.
- [6] Westfall, P. and Young, S., *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*, Wiley, 1993.
- [7] Benjamini, Y. and Hochberg, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of Royal Statistical Society Series B (Methodological)*, **57(1)**(1995), 289–300.
- [8] Balding, D.J., A tutorial on statistical methods for population association studies, *Nature Reviews Genetics*, **7(10)**(2006), 781–791.
- [9] Skol, A.D., Scott, L.J., Abecasis, G.R. and Boehnke, M., Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies, *Nature Genetics*, **38(2)**(2006), 209–213.
- [10] Wang, H., Thomas, D.C., Pe'er, I. and Stram, D.O., Optimal two-stage genotyping designs for genome-wide association scans, *Genetic Epidemiology*, **30(4)**(2006), 356–368.
- [11] Yu, K., Chatterjee, N., Wheeler, W., Li, Q., Wang, S., Rothman, N. and Wacholder, S., Flexible design for following up positive findings, *American Journal of Human Genetics*, **81(3)**(2007), 540–551.
- [12] Pan, D., Li, Q., Jiang, N., Liu, A. and Yu, K., Robust joint analysis allowing for model uncertainty in two-stage genetic association studies, *BMC Bioinformatics*, **12(9)**(2011), 1–7.
- [13] Devlin, B. and Roeder, K., Genomic control for association studies, *Biometrics*, **55(4)**(1999), 997–1004.
- [14] Pritchard, J.K., Stephens, M. and Donnelly, P., Inference of population structure using multilocus genotype data, *Genetics*, **155(2)**(2000), 945–959.
- [15] Pritchard, J.K., Stephens, M., Rosenberg, N.A. and Donnelly, P., Association mapping in structured populations, *American Journal of Human Genetics*, **67(1)**(2000), 170–181.
- [16] Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D., Principal components analysis corrects for stratification in genome-wide association studies, *Nature Genetics*, **38(8)**(2006), 904–909.
- [17] Li, Q. and Yu, K., Improved correction for population stratification in genome-wide association studies by identifying hidden population structures, *Genetic Epidemiology*, **32(3)**(2008), 215–226.
- [18] Epstein, M.P., Allen, A.S. and Satten, G.A., A simple and improved correction for population stratification in case-control studies, *American Journal of Human Genetics*, **80(5)**(2007), 921–930.
- [19] Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C. and Eskin, E., Variance component model to account for sample structure in genome-wide association studies, *Nature Genetics*, **42(4)**(2010), 348–354.
- [20] Devlin, B., Roeder, K. and Wasserman, L., Genomic control, a new approach to genetic-based association studies, *Theoretical Population Biology*, **60(3)**(2001), 155–166.
- [21] Armitage, P., Tests for linear trends in proportions and frequencies, *Biometrics*, **11(3)**(1955), 375–386.

- [22] Sasieni, P.D., From genotypes to genes: doubling the sample size, *Biometrics*, **53(4)**(1997), 1253–1261.
- [23] Mardia, K.V., Kent, J.T. and Bibby, J.M., *Multivariate Analysis*, New York: Academic Press, 2003.
- [24] Li, Q., Wacholder, S., Hunter, D.J., Hoover, R.N., Chanock, S., Thomas, G. and Yu, K., Genetic background comparison using distance-based regression, with applications in population stratification evaluation and adjustment, *Genetic Epidemiology*, **33(5)**(2009), 432–441.
- [25] Zheng, G., Freidlin, B., Li, Z.H. and Gastwirth, J.L., Choice of scores in trend tests for case-control studies of candidate-gene associations, *Biometrical Journal*, **45(3)**(2003), 335–348.
- [26] Zheng, G., Freidlin, B. and Gastwirth, J.L., Comparison of robust tests for genetic association using case-control studies, *IMS Lecture Notes-Monograph Series, 2nd Lehmann Symposium-Optimality*, Rojo, J., ed., Institute of Mathematical Statistics: Beachwood, OH, 2006, 253–265.
- [27] Wang, K. and Sheffield, V.C., A constrained-likelihood approach to marker-trait association studies, *American Journal of Human Genetics*, **77(5)**(2005), 768–780.
- [28] Zheng, G. and Chen, Z., Comparison of maximum statistics for hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrics*, **61(1)**(2005), 254–258.
- [29] Li, Q., Zheng, G., Liu, A., Xiong, S., Li, Z. and Yu, K., The limiting bound of Efron's W-formula for hypothesis testing when a nuisance parameter is present only under the alternative, *Journal of Statistical Planning and Inference*, **140(6)**(2010), 1610–1617.
- [30] Gastwirth, J.L., The use of maximin efficiency robust tests in combining contingency tables and survival analysis, *Journal of the American Statistical Association*, **80(390)**(1985), 380–384.
- [31] Freidlin, B., Zheng, G., Li, Z. and Gastwirth, J.L., Trend tests for case-control studies of genetic markers: power, sample size and robustness, *Human Heredity*, **53(3)**(2002), 146–152.
- [32] Li, Q., Zheng, G., Li, Z. and Yu, K., Efficient approximation of p-value of the maximum of correlated tests, with applications to genome-wide association studies, *Annals of Human Genetics*, **72(3)**(2008), 397–406.
- [33] The Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature*, **447(7145)**(2007), 661–678.
- [34] Joo, J., Kwak, M., Ahn, K. and Zheng, G., A robust genome-wide scan statistic of the Wellcome Trust Case-Control Consortium, *Biometrics*, **65(4)**(2009), 1115–1122.
- [35] Zheng, G. and Ng, H.K., Genetic model selection in two-phase analysis for case-control association studies, *Biostatistics*, **9(3)**(2008), 391–399.
- [36] Song, K. and Elston, R.C., A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies, *Statistics in Medicine*, **25(1)**(2006), 105–126.
- [37] Joo, J., Kwak, M. and Zheng, G., Improving power for testing genetic association in case-control studies by reducing the alternative space, *Biometrics*, **66(1)**(2010), 266–276.
- [38] Schaid, D.J., McDonnell, S.K., Hebring, S.J., Cunningham, J.M. and Thibodeau, S.N., Nonparametric tests of association of multiple genes with human disease, *American Journal of Human Genetics*, **76(5)**(2005), 780–793.
- [39] Zhang, H., Liu, C.T. and Wang, X., An association test for multiple traits based on the generalized Kendall's Tau, *Journal of the American Statistical Association*, **105(490)**(2010), 473–481.

- [40] Zhu, W., Jiang, Y. and Zhang, H., Nonparametric covariate-adjusted association tests based on the generalized Kendall's Tau, *Journal of the American Statistical Association*, **107(497)**(2012), 1–11.
- [41] Li, Q., Li, Z.B., Zheng, G., Gao, G. and Yu, K., Rank-based robust tests for quantitative-trait genetic association studies, *Genetic Epidemiology*, **37(4)**(2013), 358–365.
- [42] Jeffreys, H., *Theory of Probability*, Oxford: Oxford University Press, 1961.
- [43] Stephens, M. and Balding, D.J., Bayesian statistical methods for genetic association studies, *Nature Reviews Genetics*, **10(10)**(2009), 681–690.
- [44] Kass, R.E. and Raftery, A.E., Bayes factors, *Journal of the American Statistical Association*, **90(430)**(1995), 773–795.
- [45] Wakefield, J., A Bayesian measure of the probability of false discovery in genetic epidemiology studies, *American Journal of Human Genetics*, **81(2)**(2007), 208–227.
- [46] Wakefield, J., Bayes factors for genome-wide association studies: comparison with P-values, *Genetic Epidemiology*, **33(1)**(2009), 79–86.
- [47] Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T.J., Montpetit, A., Pshzhetsky, A.V., Prentki, M., Posner, B.I., Balding, D.J., Meyre, D., Polychronakos, C. and Froguel, P., A genome-wide association study identifies novel risk loci for type 2 diabetes, *Nature*, **445(7130)**(2007), 881–885.
- [48] Todd, J.A., Walker, N.M., Cooper, J.D., Smyth, D.J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S.F., Payne, F., Lowe, C.E., Szeszko, J.S., Hafler, J.P., Zeitels, L., Yang, J.H., Vella, A., Nutland, S., Stevens, H.E., Schuilenburg, H., Coleman, G., Maisuria, M., Meadows, W., Smink, L.J., Healy, B., Burren, O.S., Lam, A.A., Ovington, N.R., Allen, J., Adlem, E., Leung, H.T., Wallace, C., Howson, J.M., Guja, C., Ionescu-Tirgoviste, C., Simmonds, M.J., Heward, J.M., Gough, S.C., Dunger, D.B., Wicker, L.S. and Clayton, D.G., Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes, *Nature Genetics*, **39(7)**(2007), 857–864.
- [49] Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., Bitton, A., Dassopoulos, T., Datta, L.W., Green, T., Griffiths, A.M., Kistner, E.O., Murtha, M.T., Regueiro, M.D., Rotter, J.I., Schumm, L.P., Steinhart, A.H., Targan, S.R., Xavier, R.J., NIDDK IBD Genetics Consortium, Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J.P., de Vos, M., Vermeire, S., Louis, E., Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Cardon, L.R., Anderson, C.A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N.J., Onnie, C.M., Fisher, S.A., Marchini, J., Ghorri, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C.G., Parkes, M., Georges, M. and Daly, M.J., Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease, *Nature Genetics*, **40(8)**(2008), 955–962.
- [50] Zhang, X.J., Huang, W., Yang, S., Sun, L.D., Zhang, F.Y., Zhu, Q.X., Zhang, F.R., Zhang, C., Du, W.H., Pu, X.M., Li, H., Xiao, F.L., Wang, Z.X., Cui, Y., Hao, F., Zheng, J., Yang, X.Q., Cheng, H., He, C.D., Liu, X.M., Xu, L.M., Zheng, H.F., Zhang, S.M., Zhang, J.Z., Wang, H.Y., Cheng, Y.L., Ji, B.H., Fang, Q.Y., Li, Y.Z., Zhou, F.S., Han, J.W., Quan, C., Chen, B., Liu, J.L., Lin, D., Fan, L., Zhang, A.P., Liu, S.X., Yang, C.J., Wang, P.G., Zhou, W.M., Lin, G.S., Wu, W.D., Fan, X., Gao, M., Yang, B.Q., Lu, W.S., Zhang, Z., Zhu, K.J., Shen, S.K., Li, M., Zhang, X.Y., Cao, T.T., Ren, W., Zhang, X., He, J., Tang, X.F., Lu, S., Yang, J.Q., Zhang, L., Wang, D.N., Yuan, F., Yin, X.Y.,

- Huang, H.J., Wang, H.F., Lin, X.Y. and Liu, J.J., Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21, *Nature Genetics*, **41(2)**(2009), 205–210.
- [51] de Cid, R., Riveira-Munoz, E., Zeeuwen, P.L., Robarge, J., Liao, W., Dannhauser, E.N., Giardina, E., Stuart, P.E., Nair, R., Helms, C., Escaramís, G., Ballana, E., Martin-Ezquerria, G., den Heijer, M., Kamsteeg, M., Joosten, I., Eichler, E.E., Lazaro, C., Pujol, R.M., Armengol, L., Abecasis, G., Elder, J.T., Novelli, G., Armour, J.A., Kwok, P.Y., Bowcock, A., Schalkwijk, J. and Estivill, X., Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis, *Nature Genetics*, **41(2)**(2009), 211–215.
- [52] Sun, L.D., Cheng, H., Wang, Z.X., Zhang, A.P., Wang, P.G., Xu, J.H., Zhu, Q.X., Zhou, H.S., Ellinghaus, E., Zhang, F.R., Pu, X.M., Yang, X.Q., Zhang, J.Z., Xu, A.E., Wu, R.N., Xu, L.M., Peng, L., Helms, C.A., Ren, Y.Q., Zhang, C., Zhang, S.M., Nair, R.P., Wang, H.Y., Lin, G.S., Stuart, P.E., Fan, X., Chen, G., Tejasvi, T., Li, P., Zhu, J., Li, Z.M., Ge, H.M., Weichenthal, M., Ye, W.Z., Zhang, C., Shen, S.K., Yang, B.Q., Sun, Y.Y., Li, S.S., Lin, Y., Jiang, J.H., Li, C.T., Chen, R.X., Cheng, J., Jiang, X., Zhang, P., Song, W.M., Tang, J., Zhang, H.Q., Sun, L., Cui, J., Zhang, L.J., Tang, B., Huang, F., Qin, Q., Pei, X.P., Zhou, A.M., Shao, L.M., Liu, J.L., Zhang, F.Y., Du, W.D., Franke, A., Bowcock, A.M., Elder, J.T., Liu, J.J., Yang, S. and Zhang, X.J., Association analyses identify six new psoriasis susceptibility loci in the Chinese population, *Nature Genetics*, **42(11)**(2010), 1005–1009.
- [53] Li, T., Li, Z., Chen, P., Zhao, Q., Wang, T., Huang, K., Li, J., Li, Y., Liu, J., Zeng, Z., Feng, G., He, L. and Shi, Y., Common variants in major histocompatibility complex region and TCF4 gene are significantly associated with schizophrenia in Han Chinese, *Biological Psychiatry*, **68(7)**(2010), 671–673.
- [54] Stefansson, H., Ophoff, R.A., Steinberg, S., Andreassen, O.A., Cichon, S., Rujescu, D., Werge, T., Pietiläinen, O.P., Mors, O., Mortensen, P.B., Sigurdsson, E., Gustafsson, O., Nyegaard, M., Tuulio-Henriksson, A., Ingason, A., Hansen, T., Suvisaari, J., Lonnqvist, J., Paunio, T., Børglum, A.D., Hartmann, A., Fink-Jensen, A., Nordentoft, M., Hougaard, D., Norgaard-Pedersen, B., Böttcher, Y., Olesen, J., Breuer, R., Möller, H.J., Giegling, I., Rasmussen, H.B., Timm, S., Mattheisen, M., Bitter, I., Réthelyi, J.M., Magnusdottir, B.B., Sigmundsson, T., Olason, P., Masson, G., Gulcher, J.R., Haraldsson, M., Fossdal, R., Thorgeirsson, T.E., Thorsteinsdottir, U., Ruggeri, M., Tosato, S., Franke, B., Strengman, E., Kiemenev, L.A., Genetic Risk and Outcome in Psychosis (GROUP), Melle, I., Djurovic, S., Abramova, L., Kaleda, V., Sanjuan, J., de Frutos, R., Bramon, E., Vassos, E., Fraser, G., Ettinger, U., Picchioni, M., Walker, N., Touloupoulou, T., Need, A.C., Ge, D., Yoon, J.L., Shianna, K.V., Freimer, N.B., Cantor, R.M., Murray, R., Kong, A., Golimbet, V., Carracedo, A., Arango, C., Costas, J., Jönsson, E.G., Terenius, L., Agartz, I., Petursson, H., Nöthen, M.M., Rietschel, M., Matthews, P.M., Muglia, P., Peltonen, L., St Clair, D., Goldstein, D.B., Stefansson, K. and Collier, D.A., Common variants conferring risk of schizophrenia, *Nature*, **460(7256)**(2009), 744–747.
- [55] Yue, W.H., Wang, H.F., Sun, L.D., Tang, F.L., Liu, Z.H., Zhang, H.X., Li, W.Q., Zhang, Y.L., Zhang, Y., Ma, C.C., Du, B., Wang, L.F., Ren, Y.Q., Yang, Y.F., Hu, X.F., Wang, Y., Deng, W., Tan, L.W., Tan, Y.L., Chen, Q., Xu, G.M., Yang, G.G., Zuo, X.B., Yan, H., Ruan, Y.Y., Lu, T.L., Han, X., Ma, X.H., Wang, Y., Cai, L.W., Jin, C., Zhang, H.Y., Yan, J., Mi, W.F., Yin, X.Y., Ma, W.B., Liu, Q., Kang, L., Sun, W., Pan, C.Y., Shuang, M., Yang, F.D., Wang, C.Y., Yang, J.L., Li, K.Q., Ma, X., Li, L.J., Yu, X., Li, Q., Huang, X., Lv, L.X., Li, T., Zhao, G.P., Huang, W., Zhang,

- X.J. and Zhang, D., Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2, *Nature Genetics*, **43(12)**(2011), 1228–1284.
- [56] Donnelly, P., Progress and challenges in genome-wide association studies in humans, *Nature*, **456(7223)**(2008), 728–731.
- [57] Manolio, T.A., Genome-wide association studies and assessment of the risk of disease, *New England Journal of Medicine*, **363(2)**(2010), 166–176.
- [58] Stranger, B.E., Stahl, E.A. and Raj, T., Progress and promise of genome-wide association studies for human complex trait genetics, *Genetics*, **187(2)**(2011), 367–383.
- [59] Cirulli, E.T. and Goldstein, D.B., Uncovering the roles of rare variants in common disease through whole-genome sequencing, *Nature Reviews Genetics*, **11(6)**(2010), 415–425.
- [60] Ziegler, A., König, I.R. and Thompson, J.R., Biostatistical aspects of genome-wide association studies, *Biometrical Journal*, **50(1)**(2008), 8–28.
- [61] Li, Q., Zhang, H. and Yu, K., Approaches for evaluating rare polymorphisms in genetic association studies, *Human Heredity*, **69(4)**(2010), 219–228.
- [62] Li, Q., Pan, D., Yue, W., Gao, Y. and Yu, K., Evaluating rare variants under two-stage design, *Journal of Human Genetics*, **57(6)**(2012), 352–357.
- [63] Spielman, R.S., McGinnis, R.E. and Ewens, W.J., Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *American Journal of Human Genetics*, **52(3)**(1993), 506–516.
- [64] Ott, J., Kamatani, Y. and Lathrop, M., Family-based designs for genome-wide association studies, *Nature Reviews Genetics*, **12(7)**(2011), 465–474.
- [65] Laird, N.M. and Lange, C., The role of family-based designs in genome-wide association studies, *Statistical Science*, **24(4)**(2009), 388–397.
- [66] Bush, W.S. and Moore, J.H., Chapter 11: genome-wide association studies, *PLOS Computational Biology*, **8(12)**(2012), e1002822.
- [67] Jin, G., Lu, L., Cooney, K.A., Ray, A.M., Zuhlke, K.A., Lange, E.M., Cannon-Albright, L.A., Camp, N.J., Teerlink, C.C., Fitzgerald, L.M., Stanford, J.L., Wiley, K.E., Isaacs, S.D., Walsh, P.C., Foulkes, W.D., Giles, G.G., Hopper, J.L., Severi, G., Eeles, R., Easton, D., Kote-Jarai, Z., Guy, M., Rinckleb, A., Maier, C., Vogel, W., Cancel-Tassin, G., Egrot, C., Cussenot, O., Thibodeau, S.N., McDonnell, S.K., Schaid, D.J., Wiklund, F., Gönberg, H., Emanuelsson, M., Whittemore, A.S., Oakley-Girvan, I., Hsieh, C.L., Wahlfors, T., Tammela, T., Schleutker, J., Catalona, W.J., Zheng, S.L., Ostrander, E.A., Isaacs, W.B., Xu, J., International Consortium for Prostate Cancer Genetics, Validation of prostate cancer risk-related loci identified from genome-wide association studies using family-based association analysis: evidence from the International Consortium for Prostate Cancer Genetics (ICPCG), *Human Genetics*, **131(7)**(2012), 1095–1103.
- [68] Thompson, J.R., Attia, J. and Minelli, C., The meta-analysis of genome-wide association studies, *Briefings in Bioinformatics*, **12(3)**(2011), 259–269.
- [69] Evangelou, E. and Ioannidis, J.P., Meta-analysis methods for genome-wide association studies and beyond, *Nature Reviews Genetics*, **14(6)**(2013), 379–389.
- [70] Li, J., Horstman, B. and Chen, Y.X., Detecting epistatic effects in association studies at a genomic level based on an ensemble approach, *Bioinformatics*, **27(13)**(2011), i222–i229.

- [71] Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q. and Nordborg, M., An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations, *Nature Genetics*, **44**(7)(2012), 825–830.

## The Review of Genome-Wide Association Studies

PAN DONGDONG

(*School of Mathematics and Statistics, Yunnan University, Kunming, 650091*)

LI ZHENGBANG

(*School of Mathematics and Statistics, Central China Normal University, Wuhan, 430079*)

ZHANG WEI    LI QIZHAI

(*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190*)

This paper presents a systematic overview of the genome-wide association study. We mainly focus on the statistical methods. Some problems and challenges are also provided.

**Keywords:** Case-control design, GWAS, SNP, robust tests, Bayes factor.

**AMS Subject Classification:** 62-07, 62P10.