遗传 HEREDITAS (Beijing) 2008 年 5 月, 30(5): 543－549
ISSN 0253-9772 www.chinagene.cn

综 述

DOI: 10.3724/SP.J.1005.2008.00543

复杂疾病全基因组关联研究进展——遗传统计分析

严卫丽

新疆医科大学公共卫生学院, 乌鲁木齐 830054

摘要: 2005 年, Science 杂志首次报道了有关人类年龄相关性黄斑变性的全基因组关联研究, 此后有关肥胖、2 型糖尿病、冠心病、阿尔茨海默病等一系列复杂疾病的全基因组关联研究被陆续报道, 这一阶段被称为人类全基因组关联研究的第一次浪潮。文章分别介绍了全基因组关联研究统计分析的方法、软件和应用实例; 比较了关联分析中多重检验的 *P* 值调整方法, 包括 Bonferroni、递减的 Bonferroni 校正法、模拟运算法和控制错误发现率的方法; 还讨论了人群混杂对关联分析结果可能产生的影响及原理, 以及全基因组关联研究中控制人群混杂的方法的研究进展和应用实例。在全基因组关联研究的第一次浪潮中, 应用经典的遗传统计方法发现了许多基因-表型之间的关联并且能够对这些关联做出解释, 其中包括许多基因组中的未知基因和染色体区域。然而, 全基因组关联研究的继续发展需要进一步阐述基因组内基因之间相互作用、基因-基因之间的复杂作用网络与环境因素的相互作用在复杂疾病发生中的作用, 现有的统计分析方法肯定不能满足需要, 开发更为高级的统计分析方法势在必行。最后, 文章还给出了全基因组关联研究统计分析软件的相关网站信息。

关键词: 全基因组关联研究; 检验效能; 多重检验校正; 人群混杂; 重复

Genome-wide association study on complex diseases: genetic statistical issues

YAN Wei-Li

School of Public Health, Xinjiang Medical University, Urumqi 830054, *China*

Abstract: Since the first genome-wide association study on human age-related macular degeneration was reported by Science journal in 2005, a series of genome-wide association studies have been published on human complex diseases or traits, such as obesity, type 2 diabetes, coronary artery disease, Alzheimer's disease and so on. The study of human genetics has recently undergone a dramatic transition which is called "the first wave of genome-wide association study". Some issues in statistical analysis of genome-wide association studies were reviewed by this paper. First, statistical analysis guidelines, methods and examples for genome-wide association studies of different designs, including unrelated case-control studies, population-based studies, and family-based association studies; second, multiple testing correction of *P* values, including Bonferroni correction, step-down Bonferroni correction, permutation correction, and the correction based on false discovery rate; third, population stratification and its effect on inference of genotype-phenotype associations. The False Positive Report Probability has been successfully applied in a recent genome-wide association study on coronary artery disease to control the population stratification. Although genetic statistical methodology has been greatly developed in control of false positive associations caused by multiple testing or population stratification, it is still not sufficient to achieve the goal. Rep-

收稿日期: 2007–09–20; **修回日期:** 2008–01–28

基金项目: 教育部新世纪人才计划(编号: NECP-205-0899)、国家自然科学基金青年基金(编号: 30500419)和新疆教育厅重点科研计划 (编号: XJEDU2004I32)项目资助[Supported in part by NECP-205-0899, Chinese National Science Fund for Young Investigators (No.30500419) and Xinjiang Educational Key Scientific Programs (No.XJEDU2004I32)]

作者简介: 严卫丽(1970–), 女, 新疆乌鲁木齐市人, 博士, 教授。研究方向: 复杂疾病遗传流行病学。Tel: 0991-4362474; E-mail: yanweili01@yahoo.com.cn

licating genotype-phenotype associations is the only way to identify true association between genetic markers and common disease traits. The first wave of genome-wide association studies is producing an impressive list of unexpected associations between genes or chromosomal regions and a broad range of diseases. Traditional statistical techniques are adequate for the analysis and interpretation of these results. However, much more sophisticated methods of statistical analysis are likely to be required as we delve further into the genome in the search for networks of interacting gene variants, or interactions between gene-gene networks and environmental factors. Finally, some useful links about statistical software for genome-wide association studies were provided.

Keywords: genome-wide association study; power; multiple testing adjustment; population stratification; replication

全基因组关联研究(Genome-Wide Association Study, 或者称作 Whole Genome Association Study, GWA 研究), 简单的讲, 就是从人类全基因组范围内的序列变异(单核苷酸多态, Single Nucleotide Polymorphism, SNP)中, 筛选出那些与疾病性状关联的 SNPs。GWA 研究设计所需样本量大, 基因分型耗资巨大, 因此, 遗传统计分析的任务不仅要从几十万个 SNPs 中发现与疾病表型的关联, 同时需要严格控制由于人群混杂可能带来的假阳性, 以及因多重比较而带来的 I 类错误概率扩大等问题, 从大量的阳性结果中筛选出那些与疾病真正相关的基因组内序列变异。

GWA 研究不再是遗传学家们的梦想。自 2005 年 Science 杂志报道了第一项有关年龄相关性(视网膜)黄斑变性(Age-related Macular Degeneration) GWA 研究以来[1], 一系列有关肥胖[2~4]、2 型糖尿病[5~7]、冠心病[8]、精神分裂症以及相关表型如体重指数(Body Mass Index, BMI)[2~4]、甘油三酯[8]等的 GWA 研究被陆续报道, 与此同时, Genetic Epidemiology、Biometrics 等遗传统计相关的杂志也发表了大量关于 GWA 研究数据统计方法学方面的研究, 旨在探讨 GWA 研究的最佳研究设计方案以达到低成本、高效益地发现遗传标记与疾病之间的关联, 同时有效控制人群混杂和多重检验导致的假阳性问题等。本文主要介绍 GWA 研究在遗传统计分析方面的最新进展。

1 GWA 研究统计分析原理

如图 1 所示, GWA 研究的统计分析依据研究设计的不同可以采用不同的分析方法。

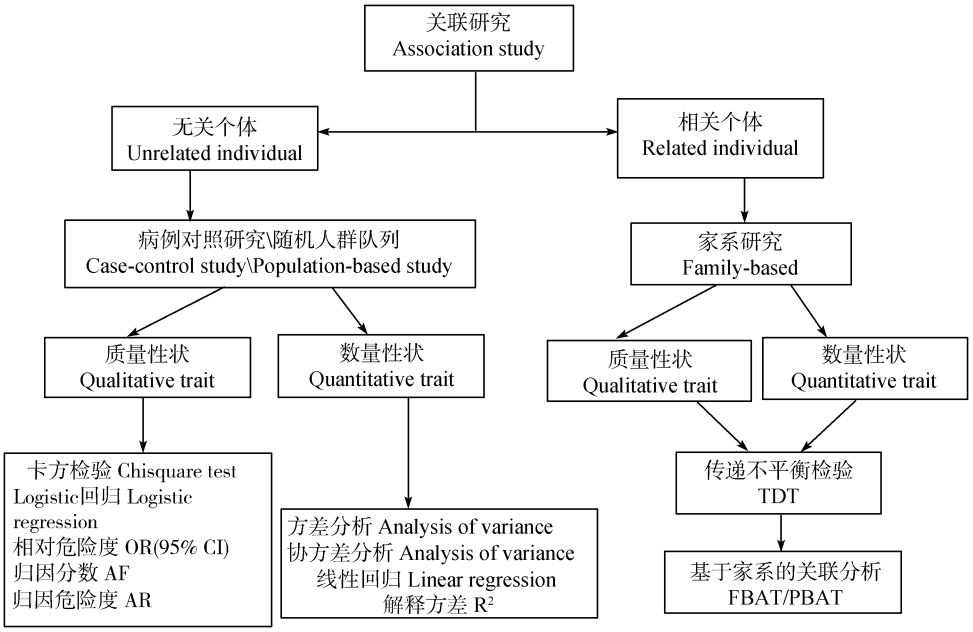

图 1 GWA 研究的统计分析方法

Fig. 1 Genetic statistical analysis of GWA study

1.1 基于无关个体(Unrelated individual)的关联分析

基于无关个体的研究设计分为病例对照研究设计(Case-control study)和基于随机人群的关联分析(Population-based association analysis)两种情况。前者主要用来研究质量性状(是否患病), 而后者主要用来研究数量性状。根据研究设计不同和研究表型的不同, 采用的统计分析方法亦不同。如病例对照研究设计(质量性状), 比较每个 SNP 的等位基因频率在病例和对照组中的差别可采用 4 格表的卡方检验, 计算相对危险度(Odds Ratio, *OR* 值)及其 95%的可信限, 进而可以计算归因分数 (Attributable fraction, *AF*)和归因危险度(Attributable risk, *AR*)。需要调整主要的混杂因素, 如年龄、性别等, 则采用 logistic 回归分析, 以研究对象患病状态为因变量, 以基因型和混杂因素作为自变量进行分析。当研究设计是基于随机人群时 (数量性状), 如研究 SNP 与某一疾病数量表型的关联时, 如 BMI, 我们比较该位点 3 种基因型携带者 BMI 水平是否有差别(单因素方差分析), 当需要调整混杂因素时, 采用协方差分析或者线性回归方程。

1.2 基于家系的关联研究(Family-based association study)

基于家系的关联研究优势之一在于可以避免人群混杂对于关联分析的影响。当研究采用家系样本时, 比如核心家系样本, 可采用传递不平衡检验(Transmisstion Disequilibrium Test, TDT)分析[9]来分析遗传标记与疾病质量表型和数量表型的关联。TDT 分析的原理是, 分析某个等位基因从杂合子的父母传递给患病孩子的机率是否高于预期值(50%)。TDT 分析的优势在于可以排除人群混杂对于关联分析的影响, 其弱点在于其发现阳性关联的检验效能低于相同样本量的病例对照研究。近年来基于家系的关联研究关联分析技术也有了明显进步。由哈佛大学和 Golden Helix INC 联合开发的 FBAT/PBAT 软件[10]是目前应用最为广泛的基于家系的统计分析工具。FBAT 是 TDT 的换代版本, 具备分析质量性状或者数量性状、调整混杂因素的作用、分析基因-环境因素的交互作用、和单体型分析的功能, 还可以对多重比较进行调整,并报告检验效能。2006 年 Science 杂志报道了一项关于肥胖的 GWA 研究结果。该研究基于 Framingham Heart Study 的家系研究样本, 以 FBAT 软件为分析工具, 先通过两阶段的基于家系的 GWA 研究, 将第一阶段全基因组范围发现的阳性关联结果按照检验效能排序, 选择检验效能最高的 10 个 SNPs 在第二阶段进行分析, 发现了 rs7566605 与肥胖的关联。随后在 5 个基于无关个体的研究样本中对所发现的关联进行重复[2]。尽管所发现的多态没有在所有的研究中得到重复, 该研究的设计, 即综合应用了 TDT 分析、两阶段研究设计、多个大样本进行重复研究多种优势于一身, 发现了用候选基因策略很难发现的遗传变异与肥胖的关联, 为复杂疾病关联研究提供了很好的经验。

以上的分析原理是基于单个位点的关联分析。关于 GWA 研究中是否有必要进行单体型分析目前尚有争论。认为有必要作单体型分析的理由是[11]: 第一, 多个位点的单体型分析有可能发现较单个位点-疾病表型之间关联更强的单体型-疾病表型之间的关联; 第二, 单体型分析有可能发现那些没有被基因分型的 SNPs (非 TagSNPs)与疾病之间的因果关联。其中第二个理由成立的条件是 GWA 研究选择了基因组内的 TagSNPs 进行基因分型。

2 GWA 研究多重假设检验调整(Multiple Testing Adjusting)

多重假设检验导致的 I 类错误扩大和假阳性关联是 GWA 研究面临的重要问题之一。多重假设检验的次数取决于所选的代表基因组的 SNPs 的数量。例如, 选择 HapMap550 则可使多重比较的次数达到 55 万次之多。有多种方法可以用来校正关联研究中多重假设检验后的 *P* 值以减少假阳性结果。图 2 是常用的几种多重比较的 *P* 值校正方法, 以及他们之间校正严格程度以及假阳性率(False positive rate)、假阴性率(False negative rate)的比较。

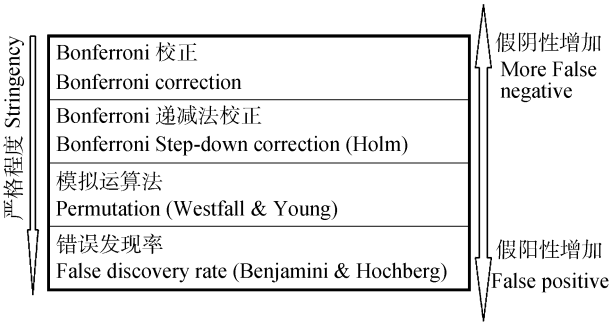

图 2　几种常见的多重检验调整方法
Fig. 2　Methods of multiple testing corrections

2.1 Bonferroni 校正法

即将单个假设检验得到的每个位点的 *P* 值乘以

本研究中同时进行假设检验的次数(即乘以所选择的遗传标记的数目)。如果校正后的 P 值仍然小于0.05, 可判断该位点与疾病之间的关联有显著性。这种校正方法被认为是多重比较 P 值调整方法中最为保守的一种方法, 存在校正过度的可能, 即增加了假阴性的机率。

2.2 递减调整法(Step-Down Adjustment)

递减调整法最早由 Holm [12] 提出, 后来由 Shaffer[13] 进行了改进。首先将所有单个位点检验获得的 P 值从小到大排序, 然后将最小的 P 值乘以所选择的位点数目 m, 排列在第二的 P 值乘以所选择的位点数目-1, 即(m-1), 其他的 P 值依次乘以 m-2, m-3, 依次类推, 排在最后的 P 值乘以 1。矫正后的 P<0.05 的位点可认为与疾病的关联有显著性。Holland[14]进一步优化了这个程序, 提高了检验效能。

以上两种方法中每个 P 值是单独进行调整的, 被称为单一步骤的校正法。下面的两种方法可以基于软件、对所有 P 值同时进行校正。

2.3 模拟运算(Permutation)法

首先对未矫正的 P 值排序[15], 然后依据基因之间结构上的关系, 通过反复抽样模拟运算, 分析 P 值的分布, 对所有的 P 值同时进行校正。

2.4 控制错误发现率(False discovery rate)法

Benjamini 和 Hochberg[16]提出了通过控制错误发现率来进行 P 值调整的方法。首先将未矫正的 P 值从小到大排序, 最大的 P 值保持不变, 其他的 P 值依次乘以系数(位点总数/该 P 值的位次)。例如, 总共 20 个位点, 对于倒数第二位的 P 值所乘系数为 20/19。以此类推。如果矫正后的 P 值<0.05, 可认为该位点与疾病的关联有显著性。相对前面 3 种校正方法而言, 这是最为宽松的一种校正方法, 因而允许更多的假阳性存在, 但同时减少了假阴性。

尽管 Bonfferoni 方法校正 P 值是最为保守和严格的一种方法, 但是为了尽可能地降低假阳性关联的概率, 目前被公认的 GWA 研究基因组水平的显著性水平 $\alpha=5\times10^{-7}$ 就是根据所选基因组 SNP 的个数采用 Bonferroni 法校正得到的。研究者们认识到, 仅仅通过校正 P 值, 无论是采用 Bonferroni 校正法, 还是模拟运算等其他方法, 都不能从根本上解决由于多重比较可能带来的假阳性关联。从 GWA 研究发现的阳性关联中发现真正的与疾病的关联, 唯一的办法只能是重复研究[17]。

3 人群混杂(Population Stratification)

人群混杂(有的文献译为人群分层)的问题对于复杂疾病的关联研究由来已久[18], 即使是基于同源性较好的同一种族研究人群也仍然存在。现有的基因组对照法(Genome Control, GC), 结构关联法(Structured Association, SA), 还有主成分分析(Principal Components)都未能有效的解决这个问题[19~25]。智联腾等[26]对前两种方法进行了系统的综述, 在此不再重复。人群混杂在 GWA 研究中仍然普遍存在, 除了技术问题, 如基因分型错误等导致假阳性关联外, 人群混杂是许多大样本研究出现假阳性、假阴性结果的重要原因之一。其原理可由表 1 和表 2 看出。Campbell 等[27]报道的乳糖酶基因(Lactase gene, *LCT*)与高个/矮个表型之间的假阳性关联就是受了人群混杂的影响。该研究所采用的研究人群为欧洲裔美国人, 其表型和基因型的关联从欧洲西北到东南方存在相当大的差异, 导致结果在其他人群不能得到重复。

与候选基因策略的关联研究相比 GWA 研究有其特殊性: 由于选择的基因组内的 SNP 数量非常大(可达 100 万个), 如果采用最为严格 Bonferroni 方法校正 P 值以校正多重比较可能带来的影响, 则有可能错过一些与疾病真正关联的 SNPs。我们在前一篇综述[28]中介绍的 GWA 研究的两阶段研究设计有两个优势: 一方面减少了基因分型的工作量和花费, 另一方面是通过重复来降低研究的假阳性率。然而

表 1 人群混杂导致的假阳性

Table 1 False positive caused by population stratification

	等位基因 *A* Allele *A*	等位基因 *B* Allele *B*	总计 Total
患病 Affected	64	16	80
未患病 Unaffected	16	4	20
总计 Total	80	20	

OR=1.0(*CL* 0.29–3.4), *P*=1

+

	等位基因 *A* Allele *A*	等位基因 *B* Allele *B*	总计 Total
患病 Affected	4	16	20
未患病 Unaffected	16	64	80
总计 Total	20	80	

OR=1.0(*CL* 0.29–3.4), *P*=1

=

	等位基因 *A* Allele *A*	等位基因 *B* Allele *B*	总计 Total
患病 Affected	68	32	100
未患病 Unaffected	32	68	100
总计 Total	100	100	

OR=4.5(*CL* 2.5–8.2), *P*=6.6 × 10^(–7)

表 2　人群混杂导致的假阴性

Table 2　False negative caused by population stratification

	等位基因 *A* Allele *A*	等位基因 *B* Allele *B*	总计 Total
患病 Affected	20	80	100
未患病 Unaffected	80	20	100
总计 Total	100	100	

OR=0.06(*CI* 0.03-0.11), *P*=4.4 × 10^(−14)

+

	等位基因 *A* Allele *A*	等位基因 *B* Allele *B*	总计 Total
患病 Affected	80	20	100
未患病 Unaffected	20	80	100
总计 Total	100	100	

OR=16.0(*CI* 8.0-31.9), *P*=4.4 × 10^(−14)

	等位基因 *A* Allele *A*	等位基因 *B* Allele *B*	总计 Total
患病 Affected	100	100	200
未患病 Unaffected	100	100	200
总计 Total	200	200	400

OR=1.0(*CI* 0.67–1.47), *P*=1

这种两阶段的研究设计很难回避这样一个棘手的、两难的问题: 第一阶段通常在较小样本中进行, 因而没有足够的检验效能发现所有可能与所研究疾病关联的 SNPs, 即假阴性。解决办法之一是适当放宽第一阶段筛选 SNPs 的标准, 将所有结果按 *P* 值排序, 根据研究的目的和财力多选择一些甚至选择一些阴性的 SNPs($P>5\times10^{-7}$), 以保证最大可能地选出与疾病关联的 SNP 进入第二阶段研究。问题是, 到底选择多少 SNPs 呢? 而且这样做同时又增加了假阳性的可能。为控制研究的假阳性概率, 有人尝试在第一阶段研究筛选 SNP 的过程中计算假阳性报告概率(False Positive Report Probability, FPRP) , 从而保证从 GWA 研究第一阶段关联分析结果中选择最优的 SNPs 进入后续阶段的研究[29]。该方法必须基于这样一个假设: 遗传标记是否与疾病关联的可能性并不取决于该标记的遗传环境(Genetic Context in Genome)无关, 比如内含子区、编码区的非同义变异(Nonsynonymous SNPs), 而且有同等的概率与所研究的疾病关联, 同时所有与疾病关联的 SNPs 对疾病发生作用大小相同, 基因频率则是影响假阳性报告概率的唯一因素。其优点在于可以估计某位点与疾病的关联为假阳性的概率大小, 被 Samani 等[8]运用, 成功地发现了与人类冠心病关联的 SNPs。此外, 来自 Emery 大学人类遗传系和美国疾病控制中心的研究者们发明了另一种新方法——分层分数法(Stratification-score approach)用来控制人群分层可能对关联分析结果可能带来的影响[30]。简言之, 该方法在第一阶段研究中用基因组内亚结构信息位点(Substructure-informative loci)分析与疾病的风险度(Odds), 不包括待测位点, 为每一个研究对象计算分层分数; 第二阶段研究将研究对象根据分层分数分层, 然后分别在每一层内分别分析待测位点与疾病的关联。除了运用统计分析的手段控制人群混杂的影响外, 采用基于家系的关联研究可以避免人群混杂对关联分析结果的影响。

4　GWA 研究的重复(Replication)

对于 GWA 研究而言, 无论是两阶段研究设计、还是各种遗传统计方法, 都无法从根本上解决由于多重比较、人群混杂等带来的假阳性问题。GWA 研究中不能单靠调整的 *P* 值水平来判断一个 SNP 是否与疾病真正关联; 即使通过不同的遗传统计分析方法也不能完全避免人群混杂导致的假阳性问题; 重复研究才是确保我们发现与疾病真关联的必要保证[17]。当一项研究发现的关联未能得到重复时, 理论上有 3 种可能: 第一、一个真的假阳性关联正确地没有被重复; 第二、一个真的关联在检测效能较低的后续研究中(比如, 样本量较小或者病例对照匹配不好)没有得到重复(假阴性); 第三、一个真的关联在一个人群里存在, 由于复杂疾病的遗传异质性, 即由于遗传和环境因素的作用而导致关联在另一个人群里不能被重复。因此, 高质量的重复研究能够帮助我们做出正确的判断。重复研究的方法之一是直接将两个或者几个研究人群联合起来进行分析, 通过增大样本量来提高研究的检验效能, 提高可能与疾病关联的 SNPs 的概率。Skol 等[31]通过模拟运算证明这种合并样本的方法比两阶段重复的方法有更大的效能, 同时对控制假阴性也较有优势。目前较受推崇的另一种重复研究方法是, 两个研究人群第一阶段研究分别在所有的研究样本中对所有的 SNPs 进行分析, 第二阶段则互相重复测量对方研究发现的阳性 SNPs。这样做的优点是, 第一阶段的研究尽可能保证了检验效能和较低的假阴性率, 第二阶段的 SNP 检测的工作量减小而样本量足够大, 同时第一阶段的阳性结果在两个独立的大样本人群进行重复, 这样就最大限度控制了研究的假阴性和假阳性率。Samani 等[8]在冠状动脉疾病的 GWA 研究中综合运用了上述几种方法来实现重复和控制假阳性, 得到较为理想的结果。他们的做法是, 先用互相重复的方法比较了英国的样本和在法国进行的相同

研究，同时也尝试了将两个样本联合起来分析的方法，而且在两阶段研究中运用了 FPRP 的方法从第一阶段筛选 SNPs 进入第二阶段研究。

要想实现大样本研究的相互重复，数据共享是必由之路。在这方面，美国国立肿瘤研究所的肿瘤遗传易感性标记(Cancer Genetic Markers of Susceptibility)项目首先做出了表率。他们将关于乳腺癌和前列腺癌的 GWA 研究的结果，包括 p 值、相对危险度(RR 值)以及 95% 可信限在网上公开(http://cgems.cancer.gov)。此外，Diabetes Genetics Initiative 的研究者们也公开了他们的研究数据(www.broad.mit.edu/diabetes)。美国国立卫生研究院制定了相关的政策督促更多的研究组共享研究数据(http://www.genome.gov/)。数据共享的重要性、必要性和优越性为越来越多的研究者们所认识和重视，这必将为 GWA 研究在人类复杂疾病的研究中得到更多有价值的发现铺平道路。

GWA 研究为复杂疾病的研究打开了新的篇章：研究者们无需像候选基因策略那样需要预先假设致病基因，而是在病例和对照中比较全基因组范围内所有变异的等位基因频率，从中发现与疾病关联的序列变异。GWA 研究已经发现了许多我们以前从未了解的未知基因和染色体区域，为我们了解人类复杂疾病的发病机制提供了更多的线索。然而，对 GWA 在复杂疾病病因研究中的作用我们也不能过于乐观。要想发现真正与复杂疾病关联的 SNPs，GWA 研究需要同时具备以下条件：GWA 研究样本的病例组必须携带着导致疾病发生的遗传因素；研究要达到足够的检验效能，所需的样本量和 SNPs 的数目都非常大，需要在几千个病例和对照中对大量的人类 SNPs 进行基因分型，数据分析也需要更为先进的统计方法。有时即使 GWA 研究结果提示了一个与疾病关联的染色体区域，进一步确定真正致病的 SNP 仍然难度很大。原因是导致疾病发生的功能 SNP 在基因内的位置变异度很大，可以在编码区、剪切位点，也可以在基因的调控区。即使 GWA 研究发现的与疾病的关联是真实的，目前只有少数人认为这些结果可以很快用于指导临床，比如用来评估一个人患某种疾病的危险。尽管 GWA 研究发现了很多可能与疾病表型相关的 SNPs，但是这些基因如何与环境因素相互作用、生活方式的改变如何调节这些基因的作用，目前仍然不清楚。伊利诺斯州芝加哥大学的 Nancy Cox 教授认为，“(GWA 研究的结果)到达临床应用的道路上仍然布满荆棘”[32]。

总之，我们正在经历人类 GWA 研究的第一次浪潮。它取得的成果是可喜的。研究者们运用传统的统计方法和技术分析和解释这些研究，发现了许多以前未知的与一系列人类复杂疾病相关的基因和染色体区域。然而，对于复杂疾病而言，这些认识还远远不够。GWA 研究还需要进一步阐明在全基因组范围内，基因-基因之间、基因-环境因素之间复杂的相互作用如何导致复杂疾病的发生。为了达到这个目标，传统的统计方法学还有待于进一步发展，还有很大的研究空间。

下面列举几个目前使用的用于 GWA 研究分析的软件和网址：

PLINK: http://pngu.mgh.harvard.edu/~purcell/plink/

FBAT/PBAT: http://www.biostat.harvard.edu/~clange/default.htm

GoldenHelix: http://www.biostat.harvard.edu/~clange/default.htm

Affymatrix, Application note, Solutions for Each step of an association study: http://www.affymetrix.com/support/technical/appnotes/assoc_study_appnote.pdf

参考文献(References):

[1] Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. *Science*, 2005, 308(5720): 385–389. [DOI]

[2] Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, Colditz G, Hinney A, Hebebrand J, Koberwitz K, Zhu X, Cooper R, Ardlie K, Lyon H, Hirschhorn JN, Laird NM, Lenburg ME, Lange C, Christman MF. A common genetic variant is associated with adult and childhood obesity. *Science*, 2006, 312(5771): 279–283. [DOI]

[3] Rosskopf D, Bornhorst A, Rimmbach C, Schwahn C, Kayser A, Kruger A, Tessmann G, Geissler I, Kroemer HK, Volzke H. Comment on “A common genetic variant is associated with adult and childhood obesity”. *Science*, 2007, 315(5809): 187: author reply 187. [DOI]

[4] Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI. A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 2007, 316(5826): 889–894. [DOI]

[5] Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC,

Meyer J, Ardlie K, Bengtsson Bostrom K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 2007, 316(5829): 1331–1336. [DOI]

[6] Ubeda M, Rukstalis JM, Habener JF. Inhibition of cyclin-dependent kinase 5 activity protects pancreatic beta cells from glucotoxicity. *J Biol Chem*, 2006, 281(39): 28858–28864. [DOI]

[7] Foley AC, Mercola M. Heart induction by Wnt antagonists depends on the homeodomain transcription factor Hex. *Genes Dev*, 2005, 19(3): 387–396. [DOI]

[8] Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, Konig IR, Stevens SE, Szymczak S, Tregouet DA, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Braenne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H. Genomewide association analysis of coronary artery disease. *N Engl J Med*, 2007, 357(5): 443–453. [DOI]

[9] Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*, 1993, 52(3): 506–516.

[10] Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, Demeo DL, Murphy A, Su J, Datta S, Rosenow C, Christman M, Silverman EK, Laird NM, Weiss ST, Lange C. Genomic screening and replication using the same data set in family-based association testing. *Nat Genet*, 2005, 37(7): 683–691. [DOI]

[11] Newton-Cheh C, Hirschhorn JN. Genetic association studies of complex traits: design and analysis issues. *Mutat Res*, 2005, 573(1-2): 54–69.

[12] Holm S. A simple sequentially rejective multiple test procedure. *Scand J Statist*, 1979, 6: 65–70.

[13] Shaffer J. Modified sequentially rejective multiple test procedures. *J Am Stat Assoc*, 1986, 81: 826–831. [DOI]

[14] Holland BS, Copenhaver MD. An improved sequentially rejective bonferroni test procedure. *Biometrics*, 1987, 43(2): 417–423. [DOI]

[15] Westfall P, Young S. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. New York: 1993.

[16] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser C Appl Stat*, 1995, 57(1): 289–300.

[17] Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni JF Jr., Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS. Replicating genotype-phenotype associations. *Nature*, 2007, 447(7145): 655–660. [DOI]

[18] YAN Wei-Li, GU Dong-Feng. Issues on association studies on complex disease. *Acta Genetica Sinica*, 2004, 31(5): 533–537.
严卫丽, 顾东风. 复杂疾病关联研究若干问题. 遗传学报, 2004, 31(5): 533–537.

[19] Devlin B, Roeder K. Genomic control for association studies. *Biometrics*, 1999, 55(4): 997–1004. [DOI]

[20] Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol*, 2001, 60(3): 155–166. [DOI]

[21] Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet*, 1999, 65(1): 220–228. [DOI]

[22] Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet*, 2000, 67(1): 170–181. [DOI]

[23] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*, 2000, 155(2): 945–959.

[24] Zhang S, Zhu X, Zhao H. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genet Epidemiol*, 2003, 24(1): 44–56. [DOI]

[25] Chen HS, Zhu X, Zhao H, Zhang S. Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Ann Hum Genet*, 2003, 67(Pt 3): 250–264. [DOI]

[26] ZHI Lian-Teng, ZHOU Gang-Qiao, HE Fu-Chu. Detection and controlling for population stratification in association studies of human complex disease. *Hereditas*(*Beijing*), 2007, 29: 3–7.
智联滕, 周钢桥, 贺福初. 人类复杂疾病关联研究中群体分层的检出和校正. 遗传, 2007, 29(1): 3–7.

[27] Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN. Demonstrating stratification in a European American population. *Nat Genet*, 2005, 37(8): 868–872. [DOI]

[28] YAN Wei-Li. Genome-wide association study on complex diseases: study design and genetic markers. *Hereditas* (*Beijing*), 2008, 30(4): 400–406.
严卫丽. 复杂疾病全基因组关联研究进展——研究设计和遗传标记. 遗传, 2008, 30(4): 400–406.

[29] Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*, 2004, 96(6): 434–442.

[30] Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet*, 2007, 80(5): 921–930. [DOI]

[31] Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*, 2006, 38(2): 209–213. [DOI]

[32] Christensen K, Murray JC. What genome-wide association studies can do for medicine. *N Engl J Med*, 2007, 356(11): 1094–1097. [DOI]