

DOI: 10.3724/SP.J.1005.2013.01065

基因水平的关联分析方法

罗旭红, 刘志芳, 董长征

宁波大学医学院预防医学系, 宁波 315211

摘要: 全基因组关联研究(Genome wide association study, GWAS)已经在国内外的医学遗传学研究中得到广泛应用,但是 GWAS 数据中所蕴含的与多基因复杂性状疾病机制相关的丰富信息尚未得到深度挖掘。近年来,研究者采用生物网络分析和生物通路分析等生物信息学和生物统计学手段分析 GWAS 数据,并探索潜在的疾病机制。生物网络分析和生物通路分析主要是以基因为单位进行的,因此必须在分析前将基因上全部或者部分单个核苷酸多态性(Single nucleotide polymorphism, SNP)的遗传关联结果综合起来,即基因水平的关联分析。基因水平的关联分析需要考虑单个 SNP 的遗传关联、基因上 SNP 数量和 SNP 之间的连锁不平衡结构等多种因素,因此不仅在遗传学的概念上也在统计方法方面具有一定的复杂性和挑战性。文章对基因水平的关联分析的研究进展、原理和应用进行了综述。

关键词: 基因水平的关联分析; 全基因组关联研究; 网络分析; 最显著 SNP 法

Advances on gene-based association analysis

LUO Xu-Hong, LIU Zhi-Fang, DONG Chang-Zheng

Department of Preventive Medicine, School of Medicine, Ningbo University, Ningbo 315211, China

Abstract: Genome-wide association studies (GWAS) have been widely used for hunting the susceptibility genes for common diseases in the past years; however, the abundant information for the disease mechanism based on the GWAS data has not been fully mined. Recently, some researchers focused on the biological network and pathway analysis for the GWAS data to explore the potential disease mechanism. Since genes are the basic units for the biological network and pathway, the genetic effects from all or part of single nucleotide polymorphisms (SNPs) in the genes should be integrated into genetic scores, which are so-called “gene-based association analysis”. Gene-based association analysis takes into account some important factors such as genetic effects of the SNPs, the number of the SNPs in the genes and the linkage disequilibrium structure of the SNPs. In this review, we will focus on the progress, principle and application of gene-based association analysis.

Keywords: gene-based association analysis; genome wide association study; network analysis; best SNP method

收稿日期: 2013-03-15; 修回日期: 2013-06-17

基金项目: 浙江省自然科学基金项目(编号: Y2100240), 宁波市自然科学基金项目(编号: 2009A610142), 浙江省卫生厅基金项目(编号: 2009A183)和宁波大学胡岚优秀博士基金资助

作者简介: 罗旭红, 硕士研究生, 专业方向: 全基因组关联分析方法。E-mail: luoxuhong@yahoo.cn

通讯作者: 董长征, 博士, 讲师, 研究方向: 生物信息学和统计遗传学。E-mail: dongchangzheng@nbu.edu.cn

网络出版时间: 2013-7-29 19:19:05

URL: <http://www.cnki.net/kcms/detail/11.1913.R.20130729.1919.001.html>

基于单核苷酸多态性(Single nucleotide polymorphism, SNP)的全基因组关联研究(Genome wide association study, GWAS)已经在国际上医学遗传学的疾病关联研究中得到广泛应用,我国也在系统性红斑狼疮^[1]、银屑病^[2]、精神分裂症^[3]、甲亢^[4]和冠心病^[5]等许多多基因复杂性状疾病的 GWAS 中取得突破性进展。但目前的 GWAS 还主要以单个 SNP 的统计分析作为核心研究手段。单个 SNP 的分析往往只挑选了最具有统计学意义的 SNP(其假阳性也最小),舍弃了其他统计学意义相对较弱的 SNP(虽然 P 值不够小未能通过多重检验,但仍可能是假阴性),这样就遗漏了与疾病存在关联但关联效应较弱的基因(相同样本量的情况下, P 值越大,关联效应也越弱),而这种基因可能占据了易感基因的大多数;另一方面,许多不是最具有统计学意义的基因上可能有多个相对独立的较具有统计学意义的 SNP,这种基因是否值得深入研究也是值得深思的。考虑到 GWAS 所消耗的巨额资金,因此越来越多研究者将目光放在对 GWAS 数据的深度数据挖掘和生物信息学分析上^[6~10]。生物网络分析是生物信息学处理 GWAS 数据的核心手段,通过网络分析既能从基因相互作用角度考虑各个基因对疾病的共同作用,又能从网络模块和生物通路角度帮助揭示疾病的分子机制。在生物网络分析中,基因是基本的分析单元(节点),它反映了基因上所有或者部分最重要的 SNP 对疾病的综合遗传关联结果,因此在网络分析之前,需要通过统计学方法将基因上全部或者部分单个 SNP 的遗传关联结果综合起来,即基因水平的关联分析。虽然有部分生物网络分析跳过基因水平的关联分析这个环节,直接在 SNP 水平上进行网络分析^[11,12],但这种方法不利于以基因为单位整合其他组学数据,也不利于疾病机制的解释。

本文将对目前国际上常见的基因水平的关联分析方法进行综述,为了表述方便,先对一些通用的符号和概念进行定义。假定某个基因上有 n 个 SNP,对每个 SNP 的等位基因与疾病是否存在关联进行检验(趋势性检验 Cochran-Armitage trend test 或等位基因检验 Allelic test),结果 P 值按从小到大顺序排列为 $P_1, P_2, \dots, P_n$, 相应的相对风险 OR 从大到小顺序排列为 $OR_1, OR_2, \dots, OR_n$ (因为 n 个 SNP 对应的病例对照的样本量一致,所以 P 值越小其对应的 OR 也

越大;假定所有 OR 均大于 1 为疾病危险因素,如 OR 小于 1 取倒数即可),其中有 m 个 P 值小于 0.05 为有统计学意义的 SNP, 分别为 $P_1, P_2, \dots, P_m$ 。“基因上的 SNP”一般采取宽松的定义,包括外显子(同义和非同义突变)、内含子、启动子、5'UTR 和 3'UTR 区域,由于启动子区域定义的比较模糊,因此常把基因上游 2~50 kb 的区域都作为基因区域。甚至有些研究考虑到增强子及其他调控区域的存在以及遗传连锁不平衡(Linkage disequilibrium, LD)区域一般在 500 kb 以内,因此把基因上下游 500 kb 区域都作为基因区域。

常见的基因水平上的关联分析方法包括最显著 SNP 法(Best SNP method)、组合法(Combination method)、回归分析方法(Regression analysis)、主成分分析(Principal component analysis, PCA)和傅里叶分析(Fourier analysis)、基于 Permutation 的方法等。

1 最显著 SNP 法

1.1 最显著 SNP 法

把基因中 P 值最小的 SNP 作为该基因的代表,其 P 值作为该基因关联分析的结果,即 $P = \min\{P_i\}$, $i=1,2,\dots,n$ 。此时 SNP 的 P 值最小,所以称为“Best SNP”或“Most-significant SNP”;因为 P 值对应的检验统计量也最大,所以又称为“Maximum statistic”。最显著 SNP 法具有最朴素的思想,计算最简单,应用的也最广泛。粗略的看,最显著 SNP 法假定最显著的 SNP 与疾病之间的关联程度在该基因上是最强的,而其他 SNP 要么与该 SNP 存在关联,要么与疾病之间的关联可以忽略不计,因此只需要把关联程度最强的 SNP 作为该基因的代表即可。

Wang 等^[6]以最显著 SNP 法为基础,改进了在表达谱分析中广泛使用的基因富集分析方法(Gene-set enrichment analysis, GSEA 或 Gene-set analysis, GSA),将基因水平上的关联分析结果用于分析在 BioCarta、KEGG 和 GO 等生物通路中是否富集了帕金森病相关基因。Torkamani 等^[13]把最显著 SNP 法获得的结果通过超几何分布来检测那些生物通路中富含疾病相关基因,以此来发现与 7 种常见疾病相关的生物通路。Askland 等^[14]将最显著 SNP 法应用到以探索性可视化方法(Exploratory visual analysis)

为基础的生物通路分析中, 研究双向情感障碍(Bipolar disorder)的遗传基础。Baranzini 等^[10]将基因水平的关联结果叠加到人类蛋白质相互作用网络中, 来搜索网络中与多发性硬化症相关的易感网络模块。

1.2 最大 OR 法

当基因上 n 个 SNP 对应的病例对照的样本量一致时, P 值越小其对应的 OR 也越大, 这时最大 OR 法与最显著 SNP 法完全等价; 当基因分型缺失造成样本量不一致时, 这时最显著的 SNP 其 OR 未必最大。最显著 SNP 法选择最具有统计学意义(假阳性最低)的 SNP 作为代表, 最大 OR 法选择与疾病关联程度最大(对疾病影响最大)的 SNP 作为代表。另外, 基因水平上的关联分析结果常作为全基因组网络分析的基本单元, 多个独立基因的 OR 值可以相乘, 相乘的结果还是 OR, 意义比较明确, 表示几个基因的联合风险; P 值则需要先进行负对数转换再使用, 而且最终的结果(也是 P 值)只能模糊的用“显著性”或“统计学意义”来解释。Chen 等^[15]首先用最大 OR 法计算每个基因对疾病的遗传得分, 然后用基因在 KEGG 代谢网络中的连接度来对遗传得分进行加权再进行网络分析, 使得在研究基因与疾病的关联中同时考虑了遗传关联和代谢网络中的基因相互作用。

1.3 次显著 SNP 法(Second best SNP)

次显著 SNP 法与最显著 SNP 非常相似, 只是将 P 值次小的而不是 P 值最小的 SNP 作为该基因的代表。Nam 等^[16]认为, 在 GWAS 中很多所谓“显著的”SNP 实际上是随机关联, 换句话说, 如果该基因与疾病存在真正的关联, 由于基因上一般都有很多 SNP, 那么除了最显著 SNP 还应存在次显著 SNP 以及更多的显著 SNP, 否则该 SNP 很可能是随机关联 SNP, 甚至最显著的结果还可能是基因分型错误造成的(如果分型模糊或缺失的结果在判别时具有倾向性, 而这种倾向性又非随机的分布在病例和对照中就可能造成这种假阳性结果, PLINK 中就有检测缺失是否非随机分布的功能), 因此用次显著 SNP 来代替最显著 SNP 会更加稳健和保守, 降低 GWAS 中存在的假阳性结果。Nam 等^[16]以次显著 SNP 法为基础, 改进了基因富集分析方法, 并开发了相应的基因水平的全基因组关联分析软件 GSA-SNP, 软件可从网站 <http://gsa.muldass.org> 免费下载。

1.4 Simes 法(Simes' method)

Simes 法原本常用在 GWAS 研究中用于多重检验校正, 同样可以用于基因水平的关联分析。它修正了最显著 SNP 法, 筛选经过调整后 P 值最小的 SNP, $P_s = \min\{\frac{nP_i}{i}\}$ 。

2 组合法

组合法假设基因上所有 SNP 都不相关即不存在 LD 关系, 所有的 P 值都是相互独立且均匀分布的, 可以通过组合的方法把所有 SNP 的遗传关联结果联合起来形成整个基因的遗传关联结果。常见的组合方法有 Fisher 组合法(Fisher's combination test)和截断乘积法(Truncated product method, TPM)。

2.1 Fisher 组合法

Fisher 组合法又称为 Fisher 乘积法(Fisher's product test), 是将基因上所有 SNP 的 P 值直接进行乘积的方法, 检验统计量 $X = -2 \ln(\prod_{i=1}^n p_i) = \sum_{i=1}^n -2 \ln(p_i)$ 服从自由度为 $2n$ 的 χ^2 分布。Fisher 组合法简单明了, 它是所有其他组合法的基础; 而且 Fisher 组合法的统计量具有明确的统计分布形式, 可以直接计算出相应的 P 值, 从应用角度来说也非常方便。Peng 等^[7]将 Fisher 组合法和生物通路分析方法相结合应用到多种疾病的 GWAS 数据分析中, 发现它们能探测到一些单纯 SNP 检验无法发现的具有较弱遗传关联的 SNP 和基因。

2.2 TPM

Zaykin 等^[17]在 Fisher 组合法的基础上提出了 TPM。Fisher 组合法的乘积是由该基因上所有 SNP 相乘得到的, 而 TPM 只关心 P 值小于 τ 的 SNP, 当 $\tau=1$ 时, TPM 等价于 Fisher 组合法; τ 常取 $\alpha=0.05$, 这时 TPM 只关心有统计学意义的 SNP, 其乘积为这些 SNP 的 P 值的乘积, 即 $W = \prod_{i=1}^m p_i$ 。由于 W 不服从 χ^2 分布, 所以不能采用 Fisher 组合法的方式计算 P 值; 实际上 m 服从二项分布, 所以 W 的计算是在二项分布基础上的。因为二项分布的计算比较复杂, 当 n 较大时可以用蒙特卡洛的方式来获得近似 P 值。组合法的基本假设是不存在 LD, 即各 SNP 相互独立;

当存在 LD 时, Zaykin 等^[17]利用 LD 系数(即其论文公式中的相关阵)修正了 W , 使 TPM 在 SNP 不独立时也能使用。模拟实验表明, 当与疾病关联的 SNP 数目较少时(如少于 10 个), TPM 的检验效能(即 Power, 检验效能越高, 假阴性就越低, 能够发现更多的阳性 SNP)高于 Fisher 组合法; 当数目较大时(如超过 20 个), 两者的效能相近甚至 Fisher 更高些。这与 TPM 的思路是一致的, 它把 P 值较大的那些 SNP 作为“负担”给舍弃了, 自然获得更高的检验效能, 相应付出的代价是不能使用 χ^2 分布来计算最终的 P 值, 而且结果解释也和 Fisher 有所差别。Yang 等^[18]将 TPM 应用到类风湿性关节炎的 GWAS 数据分析中, 验证了此前报道过的 3 个易感基因: *PTPN22*、*C5* 和 *IL2RB*, 并发现了 *HLA-DRA*、*BTNL2*、*C6orf10*、*NOTCH4*、*TAP2* 和 *TNXB* 等 6 个新易感基因。

除了上述方法, 组合法还包括排序截断乘积法(Rank truncated product, RTP)^[19]、伽马法(Inverse gamma method, GM)^[20]、极值分布法(Extreme-value distribution method)^[21]、累积趋势性检验统计量法(Cumulative trend test statistic)^[9]和优化加权组合法(Association test by combining optimally weighted markers, ATOM)^[22]等, 因为与上述方法相近或较为复杂, 在本文中就不再赘述。理论上, 伽马法的检验效能又超过上述方法, 但和分子进化分析中反映的各位点进化速率差异的伽马分布相类似, 伽马法的参数较难确定。

3 回归分析方法

回归分析方法是把基因上的多个 SNP 看作自变量 X_i , 表型(连续性指标或者患病状态)看作 Y , 通过线性或者 Logistic 回归建立 Y 与 X_i 之间的回归方程, 利用决定系数 R^2 (包括调整决定系数和伪决定系数)、方程检验的 P 值或似然值等统计量作为基因水平上的遗传关联结果。由于回归分析是直接对疾病与多个 SNP 的关联进行分析, 此时不需要在 SNP 水平上进行关联分析, 而是将 SNP 编码后(比如某个 SNP 的 3 种基因型 *AA*、*AT* 和 *TT* 可以按照 A 等位基因的数量编码成 $AA=2$, $AT=1$, $TT=0$)直接纳入方程进行多元回归分析。常见的回归分析方法包括线性回归(Linear regression)、Logistic 回归(Logistic regression)^[23,24]、岭回归(Ridge regression)^[25,26]和典型相关分析(Canonical correlation analysis)^[27]等。

3.1 线性回归

标准多元线性回归即 $Y=X\beta+\varepsilon$, 可应用于分析 s 个个体测量的连续表型(比如血压值、血糖浓度或者等离子体脂质浓度等)和 n 个 SNP 之间的关系。在遗传关联研究中, 经常将数量性状与多个基因上的多个 SNP 及年龄、性别和 BMI 等协变量进行线性回归, 以期获得经过校正后的对疾病有影响的 SNP, 其原理与利用线性回归进行基因水平的关联分析完全一致, 只是后者将 SNP 限定在同一个基因上。

3.2 Logistic 回归

当研究目标是二分类的疾病状态(病例/对照), 就需要对 s 个个体的状态(病例=1, 对照=0)和 n 个 SNP 之间的关系作 Logistic 回归分析, 即

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta^T X, \text{ 对模型检验的 } P \text{ 值可作}$$

为基因遗传关联结果; 由于 Logistic 回归系数的含义为 OR, 也可以将多个 SNP 的 OR 累乘起来作为基因对疾病的累积 OR。Ballard 等^[23]以 Logistic 回归为基础, 使用了两种不同的通路富集方法——二项分布检验和随机集方法, 用于分析 Genetic Analysis Workshop16 (GAW16)的类风湿性关节炎 GWAS 数据, 发现了与引起类风湿性关节炎有关联的 B 细胞通路、细胞激素和免疫反应通路以及抗原呈递通路。Herold 等^[24]以 Logistic 回归分析为基础, 开发了针对基因水平多个 SNP 的联合分析软件 INTERSNP。

3.3 岭回归

由于全基因组关联分析中每个基因上通常包含较多的 SNP, 而 SNP 在回归分析中被编码成 0、1 和 2, 由于自变量之间存在的相关性而在回归分析的过程中很容易产生多重共线性问题, 进而引起通过最大似然估计的回归系数的不稳定性, 影响了回归方程的建立, 而 SNP 数据中普遍存在的 LD 又加重了这个问题。在统计学上, 经常采用岭回归的方法来解决多重共线性问题, 即通过获得回归系数的最小有偏估计值(而不是无偏估计)来确保回归方程的稳定。

Cule 等^[25]将岭回归引入到基因水平的关联分析中, 并用类似岭迹图的 P 迹图(P -value trace plot)来获得岭回归方程的系数, 并将其应用到肺癌的模拟数据中, 发现岭回归不仅克服了多重共线性问题,

而且计算效率和检验效能更高, 发现了 2 个多元线性回归未发现的与肺癌存在关联的 SNP。Liu 等^[26]将广义 Logistic 岭回归(Generalized ridge Logistic regression, GRLR)引入到基因水平的关联分析, 模拟数据和两个 GWAS 数据集(海洛因成瘾和老年性黄斑变性)都证实了广义 Logistic 岭回归比传统的 Logistic 回归和组合法具有更高的检验效能。

3.4 典型相关分析

与线性和 Logistic 回归所不同的是, 典型相关分析是同时在多个表型 Y_i 与多个 SNP 自变量 X_i 之间建立回归方程, 以确定基因是否与多个相关表型之间存在关联; 当只有一个表型时, 典型相关分析就退化成多元线性回归。Tang 等^[27]将典型相关分析应用到包含 6 种指标(总白细胞数、中性粒细胞数、淋巴细胞数、单核细胞数、嗜酸性粒细胞数和嗜碱性粒细胞数)的白细胞性状 GWAS 数据集, 验证了典型相关分析能够用于基因水平的数量性状位点研究。

4 主成分分析和傅里叶分析

岭回归虽然能够处理回归方程中存在的多重共线性问题, 但是无法解决基因水平的关联分析中经常存在的高自由度问题(基因上可能存在众多 SNP 使得检验的自由度很高, 影响了检验效能), 当基因上存在较多 SNP 的时候(例如 20 个以上), 此时个体数与 SNP 数之比常常比要求的比例偏低, 导致检验效能较低也使得方程不稳定。主成分分析通过降维的方式将一组可能相关的自变量压缩成少数几个相互独立的新自变量(即主成分), 成功地解决了多重共线性问题并大大降低了回归方程的自由度。Wang 等^[28]将主成分分析用于 CHI3L2 表达水平的基因水平关联研究, Peng 等^[29]在 PCA 的基础上发展了 PCA-Bootstrap 可信区间检验方法, 并将此用于类风湿性关节炎和海洛因成瘾的数据分析。经典的主成分分析是基于线性回归的, 由于基因和表型之间的联系可能是非线性的, 因此统计学上又发展了核主成分分析, Gao 等^[30]将该方法引入到 GAW16 的数据分析中。傅里叶分析(Fourier analysis)是除主成分分析外的另一种处理高维相关数据的分析方法, 它利用傅里叶变换(Fourier transformation, FT)将基因上的多个 SNP 的基因分型结果转换成 FT 成分, 对这些

FT 成分进行加权评分获得检验统计量。Wang 等^[28]通过模拟数据将傅里叶分析与回归分析等方法进行比较, 发现傅里叶分析的可靠性和检验效能更高。

5 基于 Permutation 的方法

由于统计检验通常对数据分布都有要求(例如正态分布和均匀分布), 在统计量的分布类型不明确或者很难求得解的情况下, 统计学上常常采用蒙特卡罗(Monte Carlo)随机模拟的方式来获得相应的 P 值, 而 Permutation(重排法)和 Bootstrap(自举法)是其中两种常用方法。不管是 Permutation 还是 Bootstrap, 在统计学上都属于再抽样(Resample)方法。由于人们对总体的分布不太清楚, 或者很难进行更多的真实抽样, 于是可以考虑在原有样本基础上通过再抽样这样一种方式来产生更多的模拟样本。Permutation 是通过打乱个体的病例/对照属性(不改变 SNP 的基因型)的方式来实现再抽样, 而 Bootstrap 是直接改变 SNP 的基因型(不改变病例/对照属性)的方式来实现再抽样。理论和实践都证明, 再抽样的方法非常有效, 即使对有明确分布类型的数据, 也能获得非常逼近真实 P 值的结果。Liu 等^[31]发展了基于 Permutation 的基因水平关联分析方法, 模拟计算次数介于 1 000 次和 1 000 000 次以获得相对准确的 P 值。对身高 GWAS 数据的分析表明, 基于 Permutation 的方法能够很好的处理基因上的 LD 结构并且具有较高的检验效能。PLINK(<http://pngu.mgh.harvard.edu/~purcell/plink/>)实现的集合关联法(Set association)本质上也是一种 Permutation 基础上的集合分析方法^[32]。

6 展望

目前已经发展了不少基因水平的关联分析方法, 每种方法都有其独特的优势也有其不可忽略的缺陷, 尚未有特别公认的分析方法。最显著 SNP 法简单明了, 使用方便, 不需要太多的遗传学假设, 回避了如何综合基因上的多个 SNP 的效应这个难题, 因此在应用中使用的最频繁; 但其理论缺陷也很突出, 它只考虑了最具有统计学意义的 SNP 而忽视了基因上的其他 SNP, 这与基因水平的关联分析方法的出发点相背离。组合法的最大优势是其统计量的生物学意义突出, 即可以用累积 OR 这个简单概念来反

映基因上 SNP 的综合遗传关联结果, 而且常见的组合法计算过程中不需要原始基因分型数据, 只须提供各个 SNP 的 P 值; 但是组合法的基本假设是参与组合的各个 SNP 是相互独立的, 而这显然与实际情况并不相符, 虽然改进的组合法可以通过 Permutation 来克服 LD 的影响, 但改进的组合法更多被认为是基于 Permutation 的方法。线性和 Logistic 回归能够处理 SNP 之间的累加效应, 并且经过改进的岭回归能够克服多重共线性的影响, 但是对高自由度仍然很敏感; 主成分分析和傅里叶分析在统计学上属于比较理想的分析工具, 能够很好的处理具有高维度和相关性的数据, 但是其结果的遗传学解释比较抽象复杂, 不像组合法和 Logistic 回归中的“累积 OR”那么简单直观, 另外对普通遗传学研究者来说其易用性也要差些。基于 Permutation 的方法不需要多少遗传学假设, 而且能够保持 SNP 数据的 LD 结构, 还能够与其他方法进行组合, 因此一直深受研究者的关注, 但是由于其是通过 Monte Carlo 模拟来实现, 必然具有这类方法所共同的缺陷, 即高计算量、计算耗时; 而且基于 Permutation 的方法属于统计学上的再抽样方法, 对样本数据的依赖程度要高于其他方法, 特别是当样本量较小的时候。基因水平的关联分析方法目前还处在前沿的方法学探索阶段, 尚未被医学遗传学研究者所熟悉并应用, 还缺乏特别成功的应用案例。但是, 基因水平的关联分析方法具有非凡的遗传学和生物信息学的理论和实践意义, 在不久的将来必将像基因交互作用那样成为医学遗传学研究者的基本研究工具。

参考文献(References):

- [1] Han JW, Zheng HF, Cui Y, Sun LD, Ye DQ, Hu Z, Xu JH, Cai ZM, Huang W, Zhao GP, Xie HF, Fang H, Lu QJ, Xu JH, Li XP, Pan YF, Deng DQ, Zeng FQ, Ye ZZ, Zhang XY, Wang QW, Hao F, Ma L, Zuo XB, Zhou FS, Du WH, Cheng YL, Yang JQ, Shen SK, Li J, Sheng YJ, Zuo XX, Zhu WF, Gao F, Zhang PL, Guo Q, Li B, Gao M, Xiao FL, Quan C, Zhang C, Zhang Z, Zhu KJ, Li Y, Hu DY, Lu WS, Huang JL, Liu SX, Li H, Ren YQ, Wang ZX, Yang CJ, Wang PG, Zhou WM, Lv YM, Zhang AP, Zhang SQ, Lin D, Li Y, Low HQ, Shen M, Zhai ZF, Wang Y, Zhang FY, Yang S, Liu JJ, Zhang XJ. 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. [DOI](#)
- [2] Sun LD, Cheng H, Wang ZX, Zhang AP, Wang PG, Xu JH, Zhu QX, Zhou HS, Ellinghaus E, Zhang FR, Pu XM, Yang XQ, Zhang JZ, Xu A-E, Wu RN, Xu LM, Peng L, Helms CA, Ren YQ, Zhang C, Zhang SM, Nair RP, Wang HY, Lin GS, Stuart PE, Fan X, Chen G, Tejasvi T, Li P, Zhu J, Li ZM, Ge HM, Weichenthal M, Ye WZ, Zhang C, Shen SK, Yang BQ, Sun YY, Li SS, Lin Y, Jiang JH, Li CT, Chen RX, Cheng J, Jiang X, Zhang P, Song WM, Tang J, Zhang HQ, Sun L, Cui J, Zhang LJ, Tang B, Huang F, Qin Q, Pei XP, Zhou AM, Shao LM, Liu JL, Zhang FY, Du WD, Franke A, Bowcock AM, Elder JT, Liu JJ, Yang S, Zhang XJ. Association analyses identify six new psoriasis susceptibility loci in the Chinese population. *Nat Genet*, 2010, 42(11): 1005–1009. [DOI](#)
- [3] Yue WH, Wang HF, Sun LD, Tang FL, Liu ZH, Zhang HX, Li WQ, Zhang YL, Zhang Y, Ma CC, Du B, Wang LF, Ren YQ, Yang YF, Hu XF, Wang Y, Deng W, Tan LW, Tan YL, Chen Q, Xu GM, Yang GG, Zuo XB, Yan H, Ruan YY, Lu TL, Han X, Ma XH, Wang Y, Cai LW, Jin C, Zhang HY, Yan J, Mi WF, Yin XY, Ma WB, Liu Q, Kang L, Sun W, Pan CY, Shuang M, Yang FD, Wang CY, Yang JL, Li KQ, Ma X, Li LJ, Yu X, Li QZ, Huang X, Lv LX, Li T, Zhao GP, Huang W, Zhang XJ, Zhang D. Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. *Nat Genet*, 2011, 43(12): 1228–1231. [DOI](#)
- [4] Chu X, Pan CM, Zhao SX, Liang J, Gao GQ, Zhang XM, Yuan GY, Li CG, Xue LQ, Shen M, Liu W, Xie F, Yang SY, Wang HF, Shi JY, Sun WW, Du WH, Zuo CL, Shi JX, Liu BL, Guo CC, Zhan M, Gu ZH, Zhang XN, Sun F, Wang ZQ, Song ZY, Zou CY, Sun WH, Guo T, Cao HM, Ma JH, Han B, Li P, Jiang H, Huang QH, Liang L, Liu LB, Chen G, Su Q, Peng YD, Zhao JJ, Ning G, Chen Z, Chen JL, Chen SJ, Huang W, Song HD, The China Consortium for the Genetics of Autoimmune Thyroid Disease. A genome-wide association study identifies two new risk loci for Graves' disease. *Nat Genet*, 2011, 43(9): 897–901. [DOI](#)
- [5] Lu X, Wang L, Chen SF, He L, Yang XL, Shi YY, Cheng J, Zhang L, Gu CC, Huang JF, Wu TC, Ma YT, Li JX, Cao J, Chen JC, Ge DL, Fan ZJ, Li Y, Zhao LC, Li HF, Zhou XY, Chen LY, Liu DH, Chen JP, Duan XF, Hao YC, Wang LG, Lu FH, Liu ZD, Yao CL, Shen C, Pu XD, Yu L, Fang XH, Xu LH, Mu JJ, Wu XP, Zheng RP, Wu NQ, Zhao Q, Li Y, Liu XL, Wang MQ, Yu DH, Hu DS, Ji X, Guo DS, Sun DL, Wang QQ, Yang Y, Liu FC, Mao QX, Liang XH, Ji JF, Chen PP, Mo XB, Li DJ, Chai GP, Tang YD, Li XD, Du ZH, Liu XH, Dou CL, Yang ZL, Meng QJ, Wang D, Wang RP, Yang J, Schunkert H, Samani NJ, Kathiresan S, Reilly MP, Erdmann J, Peng XZ, Wu XG, Liu DP, Yang YJ, Chen RS, Qiang BQ, Gu DF. Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease. *Nat Genet*, 2012, 44(8): 890–894. [DOI](#)

- [6] Wang K, Li MY, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*, 2007, 81(6): 1278–1283. [DOI](#)
- [7] Peng G, Luo L, Siu H, Zhu Y, Hu PF, Hong SJ, Zhao JY, Zhou XD, Reveille JD, Jin L, Amos CI, Xiong MM. Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet*, 2010, 18(1): 111–117. [DOI](#)
- [8] Jia PL, Zheng SY, Long JR, Zheng W, Zhao ZM. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, 2011, 27(1): 95–102. [DOI](#)
- [9] Eleftherohorinou H, Wright V, Hoggart C, Hartikainen A L, Jarvelin M R, Balding D, Coin L, Levin M. Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PLoS ONE* 2009, 4(11): e8068. [DOI](#)
- [10] Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BM, Kappos L, GeneMSA Consortium, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum Mol Genet*, 2009, 18(11): 2078–2090. [DOI](#)
- [11] Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics*, 2008, 24(23): 2784–2785. [DOI](#)
- [12] Weng LJ, Macchiardi F, Subramanian A, Guffanti G, Potkin S, Yu ZX, Xie XH. SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics*, 2011, 12(1): 99. [DOI](#)
- [13] Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, 2008, 92(5): 265–272. [DOI](#)
- [14] Askland K, Read C, Moore J. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Hum Genet*, 2009, 125(1): 63–79. [DOI](#)
- [15] Chen LN, Zhang LC, Zhao Y, Xu LD, Shang YK, Wang Q, Li W, Wang H, Li X. Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways. *Bioinformatics*, 2009, 25(2): 237–242. [DOI](#)
- [16] Nam D, Kim J, Kim SY, Kim S. GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic Acids Res*, 2010, 38(S2): W749–W754. [DOI](#)
- [17] Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining *P*-values. *Genet Epidemiol*, 2002, 22(2): 170–185. [DOI](#)
- [18] Yang HC, Liang YJ, Chung CM, Chen JW, Pan WH. Genome-wide gene-based association study. *BMC Proc*, 2009, 3(Suppl. 7): S135. [DOI](#)
- [19] Dudbridge F, Koeleman BPC. Rank truncated product of *P*-values, with application to genomewide association scans. *Genet Epidemiol*, 2003, 25(4): 360–366. [DOI](#)
- [20] Zaykin DV, Zhivotovsky LA, Czika W, Shao SS, Wolfinger RD. Combining *P*-values in large-scale genomics experiments. *Pharm Stat*, 2007, 6(3): 217–226. [DOI](#)
- [21] Dudbridge F, Koeleman BP. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am J Hum Genet*, 2004, 75(3): 424–435. [DOI](#)
- [22] Li MY, Wang K, Grant SF, Hakonarson H, Li C. ATOM: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics*, 2009, 25(4): 497–503. [DOI](#)
- [23] Ballard DH, Apornntewan C, Lee J, Lee J, Wu Z, Zhao H. A pathway analysis applied to Genetic Analysis Workshop 16 genome-wide rheumatoid arthritis data. *BMC Proc*, 2009, 3(Suppl. 7): S91. [DOI](#)
- [24] Herold C, Steffens M, Brockschmidt FF, Baur MP, Becker T. INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics*, 2009, 25(24): 3275–3281. [DOI](#)
- [25] Cule E, Vineis P, De Iorio M. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 2011, 12(1): 372. [DOI](#)
- [26] Liu Z, Shen YY, Ott J. Multilocus association mapping using generalized ridge Logistic regression. *BMC Bioinformatics*, 2011, 12(1): 384. [DOI](#)
- [27] Tang CS, Ferreira MAR. A gene-based test of association using canonical correlation analysis. *Bioinformatics*, 2012, 28(6): 845–850. [DOI](#)
- [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. [DOI](#)
- [29] Peng Q, Zhao J, Xue F. PCA-based Bootstrap confidence interval tests for gene-disease association involving multiple SNPs. *BMC Genet*, 2010, 11(1): 6. [DOI](#)
- [30] Gao QS, He YG, Yuan ZS, Zhao JH, Zhang BB, Xue FZ. Gene-or region-based association study via kernel principal component analysis. *BMC Genet*, 2011, 12(1): 75. [DOI](#)
- [31] Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Hayward NK, Montgomery GW, Visscher PM, Martin NG, Macgregor S. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet*, 2010, 87(1): 139–145. [DOI](#)
- [32] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 2007, 81(3): 559–575. [DOI](#)