

利用 Monte-Carlo 模拟评价基因水平的关联分析方法

桂娟娟, 罗旭红, 张添方, 华启航, 董长征*

(宁波大学 医学院, 浙江 宁波 315211)

摘要: 采用 Monte-Carlo 模拟了多种参数条件下的遗传关联数据, 计算多种基因水平的关联分析方法的统计效能. 结果显示: 主成分 Logistic 回归分析(累积贡献率为 95%)和我们之前发展的基于 LD 结构和 Fisher 组合法的新方法(LD-Fisher)在多种参数条件下都具有很高的统计效能, Fisher 组合法只有在单个易感 SNP 的 P 值较小时或者存在多个 P 值不太小的易感 SNP 时表现较好. 主成分 Logistic 回归分析和 LD-Fisher 可以作为基因水平关联分析的较为理想的分析方法.

关键词: 遗传关联数据; 基因水平的关联分析; Monte-Carlo 模拟; 统计效能

中图分类号: Q51

文献标志码: A

文章编号: 1001-5132 (2015) 04-0104-05

以单核苷酸多态性(Single Nucleotide Polymorphism, SNP)为基础的全基因组关联研究(Genome Wide Association Study, GWAS)在挖掘多基因复杂性状疾病的易感基因方面已经取得了丰硕的成果, 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]等疾病的 GWAS 研究中获得突破. 传统的 GWAS 数据分析, 重心在于对芯片中所包含的每个 SNP 进行单独检验, 如 Cochran-Armitage 趋势性检验、等位基因分布检验和基因型分布检验等, 这样会产生如下一直困扰 GWAS 数据分析工作的困难和问题: (1)芯片中的 SNP 数量巨大, 目前一般都包含上百万个 SNP, 由此会导致严重的 I 型错误, 即假阳性问题, 需要通过多重检验进行校正, 而校正后又容易出现假阴性问题. (2)一个基因上可能有多个 SNP, 它们单独的边际效应都较弱, 但通过联合作用对疾病产生了较大作用, 而这种情况在单纯 SNP 分析中是难以体现的. (3)目前对 GWAS 数据进行通路和网络分析已经变得比较常规了, 而在这些分析中基因都是基本单位, 因此必须把单纯 SNP 分析结果转化成以基因为单位进行分析. (4)至少在目前这个阶段, 对 SNP 的分析结果在多数情况下还是要回到对基因作用的解

释上. 正因为如此, 也为了更深入挖掘 GWAS 数据中所蕴含的疾病相关信息, 近些年不少研究关注发展基因水平的关联分析方法及其在多基因疾病的生物网络分析中的应用^[6-11].

本研究发展了一种基于连锁不平衡(Linkage Disequilibrium, LD)结构和 Fisher 组合法的基因水平的关联分析方法^[12], 并利用 Monte-Carlo 模拟方法对其及另外 2 种常用的基因水平关联分析方法(Fisher 组合法和主成分 logistic 回归分析)进行评价, 了解几种方法各自的优缺点和适用范围.

1 材料与方法

1.1 Monte-Carlo 模拟遗传关联数据

1.1.1 Monte-Carlo 模拟方法

首先利用医学统计软件 SAS 中的 array $X(n)$ 语句生成 SNP 之间的相关阵, 即初始 LD 阵 R . 然后通过因子分析函数 Factor 获得其因子载荷阵 F , 进一步生成服从多元正态分布的随机变量, 即连续型基因型数据. 因为每个 SNP 的连续型基因型数据都服从标准正态分布, 所以可利用标准正态分布的分位数函数将连续型基因型数据转化为离散型基因型数据. 最后, 利用 SAS 的宏程序 Macro

收稿日期: 2014-04-24.

宁波大学学报(理工版)网址: <http://journalg.nbu.edu.cn/>

基金项目: 国家自然科学基金(31000594); 浙江省自然科学基金(Y2100240); 宁波市自然科学基金(2015A610190); 宁波大学学科项目(XKL141060).

第一作者: 桂娟娟(1989-), 女, 浙江安徽人, 在读硕士研究生, 主要研究方向: 生物信息学和统计遗传学. E-mail: 852042146@qq.com

*通信作者: 董长征(1978-), 男, 浙江丽水人, 博士/副教授, 主要研究方向: 生物信息学和统计遗传学. E-mail: dongchangzheng@nbu.edu.cn

功能, 循环执行模拟数据 SAS 程序即可获得批量模拟数据.

1.1.2 模拟参数设定

模拟参数主要包括基因上相互独立的单倍域的数量(分别为1个和2个)、每个单倍域中 SNP 数目(分别为5、8和10个)、每个单倍域中的相关系数(相关系数分别为0.7和0.9, 即 LD 系数 r^2 分别约为0.5和0.8, 代表相关强度中等和较强2种情况)、病例组低频等位基因频率(Minor Allele Frequency, MAF, 分别为0.1、0.3和0.5), 病例组和对照组样本量均为1000. 每个单倍域中均设置一个疾病易感 SNP, 风险等位基因的相对风险(Relative Risk, RR)为1.2, 其他 SNP 的 RR 则在1附近随机产生. 模拟次数为每种参数组合各模拟100次, 根据每种基因水平的关联分析方法对模拟数据分析的结果计算统计效能(Power), 以统计效能的大小来评价每种分析方法的优缺点和适用范围. 统计效能指的是当两总体确有差别时按检验水准、假设检验能发现其有差别(即拒绝 H_0)的能力. 在本文中指发现模拟所设定的易感 SNP 的能力, 1为100%能够发现模拟的易感 SNP.

1.2 常用的基因水平的关联分析方法

1.2.1 Fisher 组合法

Fisher 组合法是将基因上所有 SNP 的 P 值直接进行乘积的方法, 其检验统计量

$$X = -2 \ln \left(\prod_{i=1}^n p_i \right) = \sum_{i=1}^n -2 \ln(p_i),$$

服从自由度为 $2n$ 的 χ^2 分布(n 为 SNP 数目)^[7]. Fisher 组合法简单明了, 其统计量服从 χ^2 分布, 可以直接计算出相应的 P 值, 从应用的角度看非常方便. Fisher 组合法的基本假设是各 SNP 相互独立, 而实际上由于单倍域中 LD 结构的存在, 不少 SNP 之间存在相关性, 因此 Fisher 组合法的自由度大大膨胀了, 其统计效能也相应减少了.

1.2.2 基于 LD 结构和 Fisher 组合法的方法

结合单倍域的特征和 Fisher 组合法发展了基于 LD 结构和 Fisher 组合法的新方法(LD Based Fisher Combination Test, LD-Fisher)^[12]. 由于每个单倍域内部的 SNP 之间存在高度相关性, 因此可以利用最显著 SNP 的思想, 每个单倍域只选择 P 值最小的那个 SNP 作为该单倍域的代表, 即其他

SNP 之所以也表现为阳性, 主要是因为它们和这个 SNP 之间存在 LD. 然后考虑到每个单倍域之间是相互独立的, 因此可利用 Fisher 组合法将这些单倍域中代表 SNP 进行组合计算 P 值. 具体做法为: 首先利用 Cochran-Armitage 趋势性检验计算基因上每个 SNP 的 P 值, 然后利用 Gabriel 的单倍域置信区间算法给基因划分单倍域, 接着将每个单倍域中最显著的 SNP 筛选出来作为代表, 最后将 Fisher 组合法应用到这些相互独立的代表 SNP 上.

1.2.3 Logistic 和主成分 Logistic 回归分析

对个体的病例/对照状态(病例=1, 对照=0)和基因上的 SNP 之间的关系作 Logistic 回归分析^[13-14], 即 $\text{logit}(p) = \ln(p/(1-p)) = \beta^T X$, 对模型检验的 P 值可作为基因遗传关联结果. 由于每个 SNP 基因型编码成0、1和2, 相互之间很容易产生共线性问题, 因此对基因型进行 Logistic 回归分析是不合适的. 主成分 Logistic 回归分析^[15]通过降维的方式将一组可能相关的 SNP 压缩成少数几个相互独立的新自变量, 即主成分, 成功地解决了回归分析中经常存在的多重共线性问题, 并大大降低了回归方程的自由度.

下文将利用 Monte-Carlo 模拟遗传关联数据来对 Fisher 组合法、LD-Fisher 和主成分 Logistic 回归分析进行比较, 由于 PCA 在设置不同的累积贡献率时表现差别会较大, 因此分别比较了累积贡献率为85%(PCA₈₅)和95%(PCA₉₅)情况.

2 结果与分析

2.1 Monte-Carlo 模拟结果

图1为一次 Monte-Carlo 模拟结果的 LD 图谱

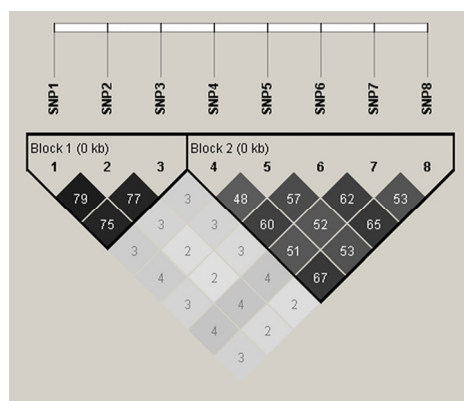


图1 2个单倍域模拟数据的 LD 图谱示例

示例. 第 1 个单倍域中有 3 个 SNP, 其 LD 系数(r^2)大小介于 0.75~0.79; 第 2 个单倍域有 5 个 SNP, 其 LD 系数(r^2)大小介于 0.48~0.67, 2 个单倍域中的 LD 系数(r^2)大小介于 0.02~0.04, 均接近于设定的参数 0.8、0.64 和 0.01.

2.2 常用基因水平的关联分析法的统计效能比较

PCA₉₅ 表现最为突出, 无论是 1 个单倍域还是 2 个单倍域的情形下, 几乎所有参数条件下其统计效能都接近 1(表 1 和表 2), 即在模拟的任何参数条件下都能捕获存在易感 SNP 的基因, 这与 PCA 原理相一致. PCA 将存在相关性的 SNP 转换成相互独立的综合变量(每个变量都综合了一个或若干个 SNP 的信息), 理论上没有损失任何信息又解决了 SNP 存在的共线性问题; 在 PCA 基础上发展的主成分 Logistic 回归分析, 因此能很好地捕获基因上所存在的疾病易感信息. 令人意外的是, PCA₈₅ 在若干模拟情况下表现得非常糟糕. 从表 1 和表 2 可发现, 在单倍域中相关系数强度为中等时(0.7), PCA₈₅ 表现得也很突出, 统计功效最低也能达到 0.85; 但当相关系数强度为较强时(0.9), PCA₈₅ 的统

表 1 1 个单倍域的模拟评价

单倍域中 SNP 数目	单倍域中 相关系数	病例组 MAF	Fisher	LD- Fisher	PCA ₈₅	PCA ₉₅
5	0.7	0.1	0.67	0.93	0.91	0.97
		0.3	0.99	1	0.99	1
		0.5	1	1	1	1
	0.9	0.1	0.61	0.95	0.12	1
		0.3	1	1	0.19	1
		0.5	1	1	0.18	1
	0.7	0.1	0.58	0.96	0.94	1
		0.3	0.97	1	1	1
		0.5	1	1	1	1
8	0.9	0.1	0.45	0.98	0.26	1
		0.3	1	1	0.15	1
		0.5	1	1	0.11	1
	0.7	0.1	0.47	0.98	0.85	1
		0.3	0.95	1	1	1
		0.5	1	1	1	1
	0.9	0.1	0.31	0.98	0.34	1
		0.3	0.96	1	0.16	1
		0.5	1	1	0.12	1

表 2 2 个单倍域的模拟评价

单倍域中 SNP 数	单倍域中 相关系数	病例组 MAF	Fisher	LD- Fisher	PCA ₈₅	PCA ₉₅
5	0.7	0.1	0.96	1	0.99	0.99
		0.3	1	1	1	1
		0.5	1	1	1	1
	0.9	0.1	0.91	1	0.15	1
		0.3	1	1	0.26	1
		0.5	1	1	0.24	1
	0.7	0.1	0.84	1	1	1
		0.3	1	1	1	1
		0.5	1	1	1	1
8	0.9	0.1	0.65	1	0.15	1
		0.3	1	1	0.09	1
		0.5	1	1	0.09	1
	0.7	0.1	0.77	1	0.99	1
		0.3	1	1	1	1
		0.5	1	1	1	1
	0.9	0.1	0.66	1	0.25	1
		0.3	1	1	0.11	1
		0.5	1	1	0.07	1

计功效都不超过 0.34. 通过对 PCA₈₅ 和 PCA₉₅ 的原始分析过程进行分析, 发现 PCA 在累积贡献率 95% 的条件下会较 85%(PCA 最常用的累积贡献率阈值)纳入更多的主成分, 其中一个主成分会偏重反映易感 SNP, 而 PCA₈₅ 因为主成分数较少, 没有单独反映易感 SNP 的主成分, 因此在 Logistic 回归分析时易感 SNP 的作用被非易感 SNP 平均化了, 这在相关系数较大时表现最为明显(这时主成分数比较少).

Fisher 在 MAF 较高时表现较为理想, 统计效能也都接近 1, 而在 MAF 较低时(0.1)表现较差, 在只有 1 个单倍域的情形下(表 1), 所有条件下的统计效能都低于 0.67; SNP 数目越多, Fisher 统计效能也越低. 这是因为 Fisher 统计量是在组合所有 SNP 的 P 值, 即 Fisher 的 P 值主要依赖于易感 SNP 的 P 值: 易感 SNP 的 P 值越小, Fisher 的 P 值也越小, 而非易感 SNP 则通过自由度的膨胀来抵消易感 SNP 的作用. 当非易感 SNP 数目越多, 自由度越大, Fisher 的 P 值也就越大; 而这种现象在 MAF 较低时表现最为明显, 因为在相同样本量情况下, 这时易感 SNP 的 P 值要明显比其他情况大. 在 2 个单

倍域的情况下(表 2), 由于增加了相互独立的易感 SNP, Fisher 的表现明显要比只有 1 个单倍域好一些, 统计效能基本都高于或等于 0.67. 当然, SNP 数目对 Fisher 的影响还是很明显, 统计效能随 SNP 数目增加表现明显的递减效应. LD-Fisher 克服了 Fisher 所存在的问题, 在只有 1 个单倍域的情形下, 统计效能介于 0.93 和 1 之间, 接近 PCA_{95} 的效能; 而在 2 个单倍域的情形下, 统计效能达到了 1, 与 PCA_{95} 的效能完全一致.

通过以上分析发现, 我们所发展的 LD-Fisher 与 PCA_{95} 在多种参数条件下都具有很高的统计效能, 几乎能够完全捕获基因上的疾病易感信息. 主成分 Logistic 回归分析在累积贡献率略低时, 如 PCA_{85} 就可能因为缺乏反映易感 SNP 的关键主成分使得统计效能急剧下降; Fisher 组合法的统计效能则依赖于易感 SNP 的 P 值, 本研究通过 RR、MAF 和样本量反映了单倍域数量和 SNP 数目等条件的综合作用.

3 讨论

基因水平的关联分析方法已经成为 GWAS 领域中研究的热点, 因为研究者希望以它为基础, 配合单纯 SNP 关联研究、基因交互作用和生物网络分析来深入的挖掘多基因疾病的易感基因和作用机制. 我们仅对较为常用的 2 种经典分析方法 (Fisher 组合法和主成分 Logistic 分析方法) 以及我们所发展的 LD-Fisher 进行 Monte-Carlo 模拟评价, 探讨它们在不同遗传参数条件下的优缺点, 以寻找各自的适用范围. 还有一些分析方法, 如上文提及的 Logistic 岭回归分析、傅里叶变换分析以及各种基于 Permutation 的方法, 因为原理复杂、编程困难并且应用较少未纳入本研究. 实际从理论上, Logistic 岭回归分析的统计效能要低于 Logistic 主成分分析, 傅里叶变换分析效能则与 Logistic 主成分分析相近, 而各种基于 Permutation 的方法往往带有各自独特的假设条件.

从模拟评价结果看, 就如同理论上所指出的, Logistic 主成分分析在处理共线性问题和综合 SNP 中的疾病相关遗传信息表现优异. 但我们意外发现, 在累积贡献率为 85% 且单倍域中的相关程度较强的情况下, PCA_{85} 因为没有纳入单独反映易感

SNP 的主成分而弱化了易感 SNP 对基因的贡献, 使得基因与疾病的关联表现为阴性. 这提示我们, Logistic 主成分分析通过主成分实现的降维作用在做基因水平的关联分析方面很容易产生错误的结果, 应该尽量避免降维, 也即采取较高的累积贡献率参数(95%). 另外需要指出的是, Logistic 主成分分析的遗传含义较为复杂抽象, 特别是在没有降维的情况下, 这也是限制主成分分析应用的重要原因. 与此相反, 传统的 Logistic 回归分析则有清晰的遗传含义(每个 SNP 的系数用来反映 OR, 基因分析的结果反映累积 OR), 当然 Logistic 回归分析又受到共线性问题的困扰; 改进的 Logistic 回归分析能够克服共线性问题, 但其遗传含义也变得复杂了. Fisher 组合法的特征非常明显, 只有在单个易感 SNP 的 P 值较小时或者存在多个 P 值不太小的易感 SNP 时表现较好, 这时其遗传含义与 Logistic 回归一样反映累积 OR. 我们所发展的 LD-Fisher 则克服了 Fisher 的缺陷, 在模拟中达到了 PCA_{95} 效能, 而且比 Fisher 还具备清晰的遗传含义: 基因上多个独立易感 SNP 的累积 OR. 当然 LD-Fisher 在理论上也受到一些影响因素的困扰, 例如, 如果 SNP 之间具备一定相关性但又达不到构成单倍域的条件时(因为模拟不太容易实现, 所以本文未进行这种参数条件的评价), LD-Fisher 无法通过单倍域来判断易感 SNP 是否相互独立, 这时其统计效能必然下降, 而 Logistic 主成分分析恰恰最擅长处理这种情况. 本文还存在其他一些缺陷, 如对基因水平的关联分析结果未进行多重检验校正, 而这在 GWAS 研究中经常需要考虑的. 实际上我们最关注的不是每种方法的具体 P 值, 而是它们在各种参数条件下的具体表现规律, 这在我们的模拟中已经能够清楚地反映了. 当然在实际应用时更加理想的处理方式是再结合保持 LD 结构的 Permutation 处理.

参考文献:

- [1] Han J W, Zheng H F, Cui Y, et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus[J]. Nat Genet, 2009, 41(11):1234-1237.
- [2] Sun L D, Cheng H, Wang Z X, et al. Association analyses identify six new psoriasis susceptibility loci in the

- Chinese population[J]. *Nat Genet*, 2010, 42(11):1005-1009.
- [3] Yue W H, Wang H F, Sun L D, et al. Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2[J]. *Nat Genet*, 2011, 43(12):1228-1231.
- [4] Chu X, Pan C M, Zhao S X, et al. A genome-wide association study identifies two new risk loci for Graves' disease[J]. *Nat Genet*, 2011, 43(9):897-901.
- [5] Lu X, Wang L, Chen S F, et al. Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease[J]. *Nat Genet*, 2012, 44(8): 890-894.
- [6] Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies[J]. *American Journal of Human Genetics*, 2007, 81(6):1278-1283.
- [7] Peng G, Luo L, Siu H, et al. Gene and pathway-based second-wave analysis of genome-wide association studies[J]. *European Journal of Human Genetics: EJHG*, 2010, 18(1):111-117.
- [8] Jia P, Zheng S, Long J, et al. dmGWAS: Dense module searching for genome-wide association studies in protein-protein interaction networks[J]. *Bioinformatics*, 2011, 27(1):95-102.
- [9] Eleftherohorinou H, Wright V, Hoggart C, et al. Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases[EB/OL]. [2013-11-30]. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0008068>.
- [10] Baranzini S E, Galwey N W, Wang J, et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis[J]. *Human Molecular Genetics*, 2009, 18(11):2078-2090.
- [11] Luo X H, Liu Z F, Dong C Z. Advances on gene-based association analysis[J]. *Hereditas (Beijing)*, 2013, 35(9): 1065-1071.
- [12] Duan S, Luo X, Dong C. Identification of susceptibility modules for coronary artery disease using a genome wide integrated network analysis[J]. *Gene*, 2013, 531(2):347-354.
- [13] Ballard D H, Apornetewan C, Lee J, et al. A pathway analysis applied to genetic analysis workshop 16 genome-wide rheumatoid arthritis data[J]. *BMC Proceedings*, 2009, 3(Sup):91.
- [14] Herold C, Steffens M, Brockschmidt F F, et al. INTERSNP: Genome-wide interaction analysis guided by a priori information[J]. *Bioinformatics*, 2009, 25(24): 3275-3281.
- [15] Wang K, Abbott D. A principal components regression approach to multilocus genetic association studies[J]. *Genetic Epidemiology*, 2008, 32(2):108-118.

Evaluation of Gene Based Association Tests by Monte-Carlo Simulation

GUI Juan-juan, LUO Xu-hong, ZHANG Tian-fang, HUA Qi-hang, DONG Chang-zheng*

(School of Medicine, Ningbo University, Ningbo 315211, China)

Abstract: Monte-Carlo is adopted to evaluate the power of multiple gene based association tests. With PCA logistic regression (95% cumulative variance) and our previously developed method, the LD-Fisher is found to have the highest power, and performs well under all simulated circumstances. Fisher combination test presents high power only if the P value of the susceptibility SNP takes very small value or there are multiple susceptibility SNPs. In conclusion, PCA logistic regression and LD-Fisher can be used as gene based association tests for complex diseases.

Key words: genetic association data; gene based association test; Monte-Carlo simulation; power

(责任编辑 史小丽)