应 用

基于 R 语言和 PLINK 程序的遗传性疾病的遗传位点分析

李 鹏, 刘希望 *, 张双飞, 赵 芳, 罗文志

(汕头大学 理学院, 广东 汕头 515063)

摘 要: 目前发现很多疾病与遗传相关, 而遗传性疾病的遗传位点分析尚未得到有效的研究方法. 现以 1000 个样本为例, 首先利用 R 语言将基因位点进行二进制和十进制编码, 然后用概率统计的方法和 R 语言对数据进行预处理, 接着利用 PLINK 程序对样本进行数据分析, 结合 R 语言得到较为清晰的 Manhattan 图及 Cytoscape 网络可视化模型. 首次结合 R 语言和 PLINK 程序对遗传性疾病的遗传位点分析, 能提高整个数据分析的速度, 减少数据分析的成本, 为解决遗传性疾病的遗传位点分析提供技术支持. 同时此模型可以更方便地应用于实际生活中.

关键词: R 语言; PLINK; Cytoscape; 遗传位点 (SNPs); GWAS

1 引言

随着分子生物学检测技术的不断进步, 全基因组关联研究 (GWAS) 成为研究复杂疾病的遗传致病机理的最重要的研究方法之一 [1−2]. 近年来发表的大量结果发现了与人类各种复杂性疾病或性状相关的大量基因组区域或易感基因, 为进一步揭示其发病机制奠定了基础. 人的每个染色体都带有一个携带遗传密码的 DNA 分子, DNA 是双螺旋长链分子, 由携带四种碱基 C, A, G 和 T 的脱氧核苷酸组成. 其中, 基因则是位于 DNA 长链中的片段. 在组成 DNA 的大量碱基对中, 有一些特定的位置上的核苷酸通常会产生突变, 这些位置被称为位点. 研究发现, 人体对疾病的敏感性的差异可能与特定的基因有关 [3−4]. 因此, 在基因上定位与疾病相关的位点, 有利于研究者认识某些疾病的遗传机制, 从而获得防治相应疾病的方法. 但以往分析手段落后, 且分析周期较长, 因此决定采用 PLINK, Cytoscape 手段, 揭示基因组与遗传性疾病之间的关系.

2 计算方法

现在的计算机系统大多数使用二进制系统, 同时数据主要以补码形式存储在计算机中. 由于其只使用 0、1 两个数字符号, 所以非常简单方便, 且易于用电子方式实现. 二进制具有可行性、简易性、逻辑性和可靠性. 而遗传算法采用的编码策略, 有十进制编码遗传算法和二进制编码遗传算法两种 [5−6].

收稿日期: 2018-10-28

* 通信作者

若使用十进制对染色体进行编码, 在进行交叉变异工作的时候会提高程序的速度. 同时, 交叉变异的工作只要用简单的数学运算就能完成, 这也降低了对遗传算法编程的难度. 在遗传算法中会重复对染色体进行交叉变异, 所以提高交叉变异的运行速度是非常有价值的工作[7−8]. 如表 1 为部分数值编码结果.

表 1　部分数值编码结果

rs3094315	TT	TC	TT	TT	TC	TC
rs3131972	CT	CT	TT	CC	CT	CC
rs3131969	CC	CT	CC	CC	CT	CC
rs3094315	44	42	44	44	42	42
rs3131972	24	24	44	22	24	22
rs3131969	22	24	22	22	24	22
rs3094315	1111	1101	1111	1111	1101	1101
rs3131972	0111	0111	1111	101	0111	0101
rs3131969	0101	0111	0101	0101	0111	0101

本文采用的 PLINK[9−10] 程序包是由哈佛大学的 Shaun Purcen 博士开发的用于遗传统计的开源软件. 它的优点是可以快速分析数据, 以满足常规基因相关研究的需要.

针对本文课题, 我们决定采用 PLINK 软件进行全基因组关联分析有关数据并求解. 首先利用 PLINK 软件在所有研究对象中对选中的 SNP 进行基因分析, 然后分析每个 SNP 与疾病的关联, 分别计算检验显著性值和发病风险率. 首先根据要求将原始数据 geno.dat 文件转化成 genotype.ped, 再进一步转化 phenotype.txt 文件为 geno.map 然后通过 PLINK 中的命令进行关联性分析, 命令为 plink –ped geno.ped –map geno.map –maf 0.05 –assoc 得到 plink.assoc 文件, 将其用 notepad 打开然后保存成.txt 文件, 然后用 excel 编辑进而能过 P 值进行排序. 首先根据要求将保存样本基因序列信息的文件 geno.dat 转化成二进制文件 genotype.ped, 其次将保存样本基因序列类型的 phenotype.txt 转化为二进制文件 geno.map, 然后通过 PLINK 软件进行关联性分析, 详细命令为 plink –ped geno.ped –map geno.map –maf 0.05 –assoc, 得到分析后 plink.assoc 结果文件, 将其用 notepad++ 打开保存成文本文件, 然后用 R 语言通过 P 值进行排序 [11].

P 值是用于衡量结果可靠性的递减指标 [12−14]. P 表示事件发生的概率. 在科学研究中, $P \leq 0.05$ 的结果通常被认为是具有统计显著性的. 在分析数据的关联性中, 赋给 P 一个初值, 便可以得到一些以 P 值为阈限值且具有关联性的数据. 卡方检验是一种非常通用的假设检验 [15], 它是统计样本的实际观察值与理论估计值之间的偏差程度, 卡方值越小, 说明偏差程度越小. 为了更好的说明本论文所做的工作, 我们绘制了分析流程图 1.

图 1 为数据分析流程图：首先我们对数据进行简单的分析, 其次进行质控处理, 去掉不符合的数据并进行整理, 然后进行 Plink 分析, 获得可视化的结果, 最后进行卡方检验, 考察所得结果是否符合我们的期望.

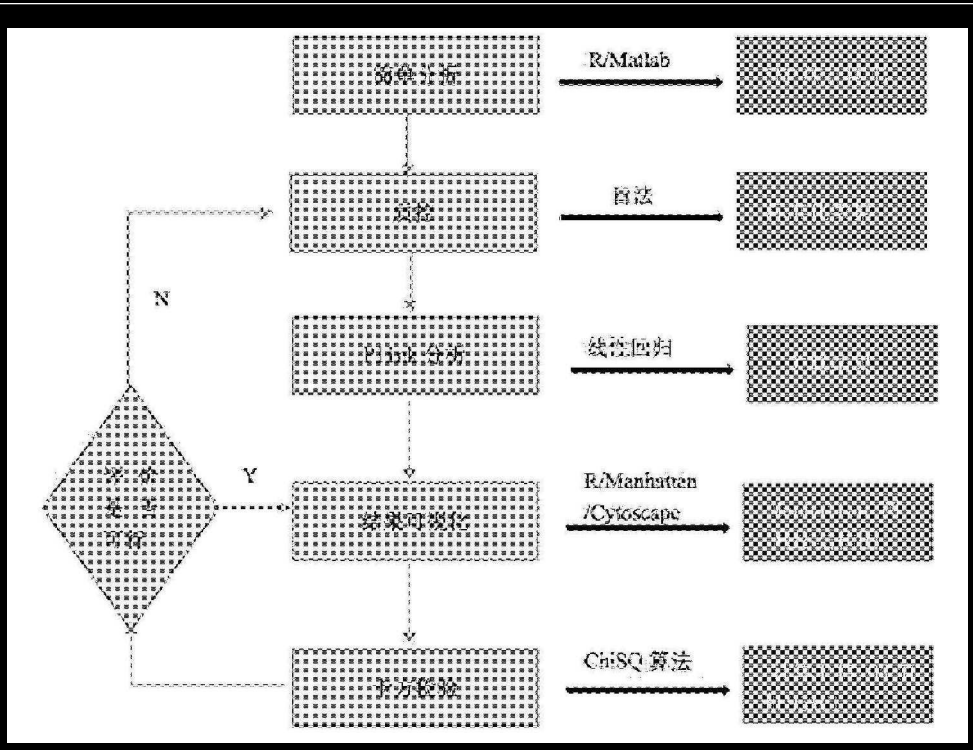

图 1 分析流程图

3 结果与讨论

3.1 数据获取

其具体方法是招募大量志愿者, 包括患有特定遗传疾病的人和健康人. 我们用 0 表示健康人, 1 表示病人. 对于每个样本, 使用碱基 (C, T, A, G) 编码来获取每个位点的信息; 例如表 2 中, 在位置 rs56341 上, 不同的编码是 C 和 A 的组合, 并且共有三种不同的编码方法: CC, CA 和 AA. 研究者可以通过将样本中的健康状况与位置编码进行比较来定位致病基因, 从而揭示遗传疾病的遗传机制.

表 2 以 2 名患者和 2 名健康人为例的 4 个样本

样本编号	健康状况	染色体片段位点名称和位点等位基因信息			
		rs100015	rs56341	···	rs21132
1	1	TT	CA	···	GT
2	0	TT	CC	···	GG
3	1	TC	CC	···	GG
4	0	TT	AA	···	GG

注: 位点名称通常以 rs 开头.

本文所引用的数据、资料均真实可靠. 针对某遗传病 A 提供了 1000 个样本的信息, 其为官方专业数据集, 且其中 500 名健康者和 500 名患病者依次在相同位点取值作为样本, 选取的每个试验者的基因片段上都有 9445 个位点 (数据来源: 全国研究生数学建模竞赛).

3.2　致病位点结果分析

对数据用 Plink 进行处理, 然后使用 R 语言中 Manhattan() 函数, 形成可视化的结果, 如图 2 所示.

横坐标为某染色体按物理图谱顺序排列的标记位点. 纵坐标为关联分析 P 值 —log10 结果, 虚线为基因组水平 1%显著的阈值线.

使用 PLINK 分析数据后, 结果输入 Cytoscape 作图得到某疾病 A 与致病位点的网络结构, 图 3 中边点与源点越近, 表示关联度越高

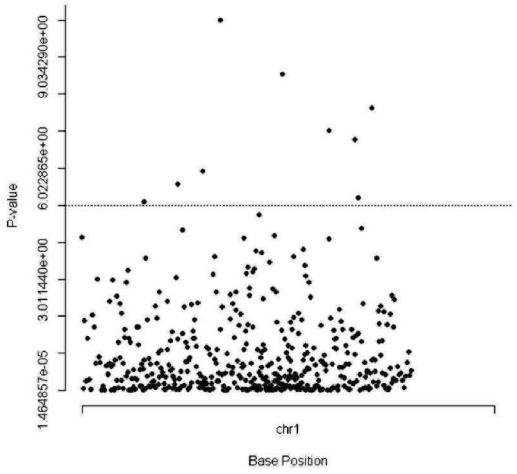

图 2　GWAS 分析曼哈顿图

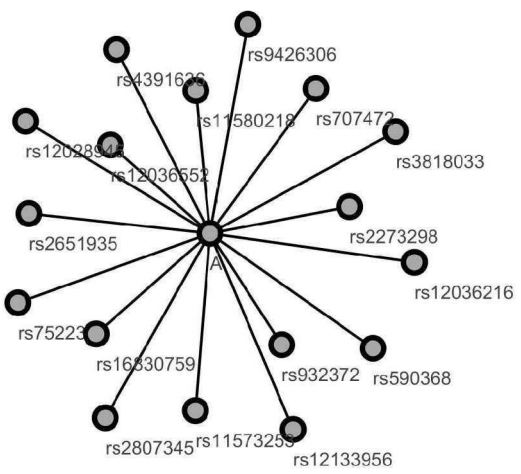

图 3　疾病 A 与致病位点的网络结构图

表 3　疾病 A 性状全基因关联分析结果中的达到 1%基因组水平显著的位点信息

性状	位点 (SNP)	发病风险率 (OR)	位置/bp	P 值	卡方检验
疾病 A	rs2273298	1.7190	2938	6.38E-08	29.25
疾病 A	rs932372	1.7420	7737	8.03E-05	15.55
疾病 A	rs12036216	0.5810	80	9.57E-05	15.22
疾病 A	rs2807345	1.5000	6794	1.74E-04	14.09
疾病 A	rs4391636	0.7005	962	2.41E-04	13.48
疾病 A	rs7522344	1.3940	1593	2.50E-04	13.41
疾病 A	rs9426306	1.3920	8589	2.58E-04	13.35
疾病 A	rs12133956	0.7027	478	6.02E-04	11.77
疾病 A	rs11580218	0.7065	6841	9.64E-04	10.90

由图 2 可知, 共有 9 个 SNPs 位点达到 1%显著基因组水平, 说明这九个位点有可能为疾病 A 的致病位点. 达到 1%显著基因组水平的具体位点信息见表 3, 显示了 500 个病例和 500 个对照样本中 SNPs 位点的所在染色体位置、P 值、位置、卡方检验以及 OR 值等与疾病 A 性状之间关系的统计数据. 其中 OR(odds ratio; 优势比) 是某种推测的概率比其反向推测的

概率大多少. 本文中其代表一种发病风险率, 文中利用 PLINK 软件分析每个 SNP 与疾病 A 的关联, 计算出发病风险率 (OR). 从表 3 中可以发现最有可能的致病位点为 rs2273298. 同时从图 3 中我们也可以发现位点 rs2273298 与疾病 A 的关联性最高.

3.3 模型优点

1) 传统的计算需要复杂的编程语言且尚未得到有效的解决, 而采用 PLINK 分析方法, 操作起来较为简单方便, 且对本分析只需要 1h 时间;

2) 采用生物信息学工具, 包括 R 语言和 Cytoscape 软件等, 可以简单编程完成;

3) 首次把 PLINK 方法与 Cytoscape 网络结构图结合起来, 发现可以更好地解释问题;

4) 此模型可以更容易应用于实际生活中, 且更直观的反映结果.

4 总结

本文通过全基因组关联信息分析 (GWAS) 以及 R 语言和 PLINK 程序进行数据处理与分析, 发现了与疾病 A 性状的基因组水平显著关联位点有 rs2273298 等 9 个 SNPs, 因生物医学上与遗传病相关的基因往往只有一个位点. 所以我们推测 rs2273298 位点为患病基因位点. 这些数据可以为医学上某些遗传疾病提供参考依据, 同时加快遗传位点在遗传性疾病上的研究. 本论文处理方法的优势是能提高整个数据分析的效率, 并减少分析成本, 为解决遗传性疾病的遗传位点分析提供了有力工具. 同时此模型也可以更方便地应用于实际生活中.

参考文献

[1] Fridley B L, Biernacka J M. Gene set analysis of SNP data: benefits, challenges, and future directions[J]. European Journal of Human Genetics, 2011, 19(8): 837-843.

[2] Yang J, Lee S H, Goddard M E, et al. GCTA: a tool for genome-wide complex trait analysis[J]. The American Journal of Human Genetics, 2011, 88(1): 76-82.

[3] 罗德威. 基于胃癌 GWAS 筛选的 SNPs 与胃癌预后的关联性研究 [D]. 南京医科大学, 2012.

[4] Lambert J C, Ibrahim-Verbaas C A, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease[J]. Nature genetics, 2013, 45(12): 1452-1458.

[5] 刘美玲, 曾德胜, 谢冲. 基于十进制编码改进的遗传算法 [J]. 广西民族大学学报 (自然科学版), 2006, 12(3): 92-94.

[6] 唐飞, 孙治国. 十进制编码遗传算法的模式定理研究 [J]. 小型微型计算机系统, 2000, 21(4): 346-367.

[7] 莫鸿强. 遗传算法搜索能力和编码方式研究 [D]. 广州: 华南理工大学图书馆, 2001.

[8] 李敏强, 寇纪淞, 林丹, 等. 遗传算法的基本理论与应用 [M]. 科学出版社, 2002.

[9] Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses[J]. The American Journal of Human Genetics, 2007, 81(3): 559-575.

[10] Chang, Christopher C, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets[M]. Gigascience, 2015, 4(1): 7.

[11] 潘东东. 全基因组关联研究中的两阶段设计与分析 [D]. 云南大学, 2012.

[12] Royall, Richard M. The effect of sample size on the meaning of significance tests[J]. The American Statistician, 1986, 40(4): 313-315.

[13] 闫懋博, 田茂再. 基于变量选择事件的新弹性网方法 [J]. 数学的实践与认识, 2019, 49(12): 215-226.

[14] 刘成友, et al. 基于基因表达谱数据筛选差异表达基因新方法 [J]. 数学的实践与认识, 2016, 46(18): 122-128.

[15] Campbell, Ian. Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations[J]. Statistics in Medicine, 2007, 26(19): 3661-3675.

Genetic Site Analysis of Hereditary Diseases Based on R-language and PLINK Program

LI Peng, LIU Xi-wang, ZHANG Shuang-fei, ZHAO Fang, LUO Wen-zhi

(College of Science, Shantou University, Shantou 515063, China)

Abstract: Many diseases have been found to be genetically related, and genetic site analysis of hereditary diseases has not been studied effectively. In the case of 1000 samples, we first use the R language to binary and decimal the gene locus. Then we use the probability statistics method and the R language to preprocess the data. Then we use the PLINK program to analyze the data, combine the R language Get a clearer Manhattan diagram and Cytoscape network visualization model. We combines R language and PLINK program for the first time to analyze the genetic site of hereditary diseases, which can increase the speed of the entire data analysis, reduce the cost of data analysis. And it can provide technical support for solving the genetic site of hereditary diseases. At the same time, this model can be more conveniently applied in real life.

Keywords: R language; PLINK; Cytoscape; Single nucleotide polymorphisms (SNPs); GWAS