遗传 HEREDITAS (Beijing) 2013 年 12 月, 35(12): 1331－1339
ISSN 0253-9772 www.chinagene.cn

综 述

DOI: 10.3724/SP.J.1005.2013.01331

# 基于单核苷酸多态性的基因互作分析方法学进展

栾奕昭[1], 左晓宇[2], 刘轲[2], 李谷[1], 饶绍奇[1,2]

1. 广东医学院医学系统生物学研究所与公共卫生学院, 东莞 523808;
2. 中山大学公共卫生学院, 广州 510080

**摘要:** 基于单核苷酸多态性的关联分析已成为当前解析人类常见复杂疾病遗传机制的重要手段之一, 然而, 目前普遍使用的单位点分析策略仅能发现部分单独效应显著的易感 SNP 位点, 因此遗漏了重要的遗传力组分——基因上位效应或联合效应。识别全基因组多基因间复杂的互作关系已成为全面解析复杂疾病致病分子机制必不可少的一项任务。已有很多方法被应用于全基因组交互作用分析, 加深了人类对复杂疾病遗传机制的进一步认识。基于各类方法的理论基础及算法的异同, 文章对目前应用较为广泛的基于遗传互作模型的方法、不基于互作模型的方法和数据挖掘类算法 3 类方法进行了系统地评述, 着重介绍了这些方法的主要思想、实现过程及应用中的注意事项等, 并指出开展大规模全基因组范围互作检测面临的问题, 以期能为相关领域的研究者提供方法学参考。

**关键词:** SNP/基因互作; 模型依赖; 数据挖掘算法; 下游功能学分析

# Advances in development of gene-gene interaction analysis methods based on SNP data: a review

LUAN Yi-Zhao[1], ZUO Xiao-Yu[2], LIU Ke[2], LI Gu[1], RAO Shao-Qi[1,2]

1. *Institute for Medical Systems Biology and School of Public Health*, *Guangdong Medical College*, *Dongguan* 523808, *China*;
2. *School of Public Health*, *Sun Yat-sen University*, *Guangzhou* 510080, *China*

**Abstract:** The SNP-based association analysis has become one of the most important approaches to interpret the underlying molecular mechanisms for human complex diseases. Nevertheless, the widely-used singe-locus analysis is only capable of capturing a small portion of susceptible SNPs with prominent marginal effects, leaving the important genetic component, epistasis or joint effects, to be undetectable. Identifying the complex interplays among multiple genes in the genome-wide context is an essential task for systematically unraveling the molecular mechanisms for complex diseases. Many approaches have been used to detect genome-wide gene-gene interactions and provided new insights into the genetic basis of complex diseases. This paper reviewed recent advances of the methods for detecting gene-gene interaction, categorized

收稿日期: 2013–06–08; 修回日期: 2013–07–24
基金项目: 国家自然科学基金项目(编号: 30830104, 31071166), 广东省科技计划攻关项目(编号: 2009A030301004), 东莞市科技重点项目(编号: 201108101015)和广东医学院基金项目(编号: XG1001, XZ1105, STIF201122)资助
作者简介: 栾奕昭, 硕士研究生, 专业方向: 流行病与卫生统计学。Tel: 13929429587; E-mail: luanyz_leo@163.com
通讯作者: 饶绍奇, 教授, 博士生导师, 研究方向: 遗传统计与生物信息学方向。E-mail: raoshaoq@gdmc.edu.cn
网络出版时间: 2013-9-12 0:57:14
URL: http://www.cnki.net/kcms/detail/11.1913.R.20130912.0057.003.html

into three types, model-based and model-free statistical methods, and data mining methods, based on their characteristics in theory and numerical algorithm. In particular, the basic principle, numerical implementation and cautions for application for each method were elucidated. In addition, this paper briefly discussed the limitations and challenges associated with detecting genome-wide epistasis, in order to provide some methodological consultancies for scientists in the related fields.

**Keywords:** SNP/gene-gene interaction; model dependence; data mining algorithm; downstream function analysis

基于单核苷酸多态性(Single nucleotide polymorphism, SNP)的关联分析已成为当前解析人类常见复杂疾病(如冠心病、糖尿病、哮喘等)遗传机制的重要手段之一。然而, 目前普遍使用的单位点关联分析策略仅能发现部分单独效应显著的易感SNP位点, 遗漏了重要的遗传力组分——基因互作效应或联合效应[1]。多数复杂疾病的发生与发展不仅仅是少数几个主效应基因的作用, 常牵涉多个微效或弱效基因的相互作用[2]。事实上, 识别多基因间复杂的互作关系已成为全面解析复杂疾病致病分子机制的重要手段之一。

近年来发展了一系列检测基因互作的统计方法或生物信息学算法。根据是否依赖互作模型的假设, 大致可分为基于互作模型的统计方法(如 logistic 回归模型等)、不基于互作模型的统计方法和数据挖掘类算法(如决策树分析、遗传规划算法等)。基于互作模型的统计方法主要以等位基因/单体型频率或其衍生测度作为分析指标, 按照预先设定的互作遗传模式剖分位点间的各种效应, 并对模型定义的互作参数进行统计学检验; 不基于互作模型的统计方法和数据挖掘类算法通常不依赖遗传模型, 通过构建统计量来检验遗传互作的存在, 或直接利用某一数据挖掘方法和计算机优化算法搜索基因互作空间中的最优解。本文重点对目前广泛应用于分类性状(如患病状态)的基因互作分析方法进行了系统地评述, 并指出了在当前实际应用过程中存在的一些问题, 以期能为相关领域的研究者提供参考。

## 1 基于互作模型的统计分析方法

基于互作模型的基因互作分析方法应用广泛, 具有良好的统计理论基础和较成熟的方法学框架。该类方法的基本思想是：根据预先设定的互作遗传模型, 剖分位点间的各项效应, 最后对模型的特定互作参数进行统计学检验。本文介绍几种常用的基于模型的两位点互作分析。

### 1.1 logistic 回归模型

logistic回归[1]是一种较为传统的用于基因-表型关联分析和基因互作分析的方法。假设有两个SNP位点$G$和$H$, 其等位基因分别为$\{G_1,G_2\}$和$\{H_1,H_2\}$, 其中$G_1$、$H_1$为风险等位基因, 基于模型的两位点logistic回归模型可定义如下：

$$\ln\left(\frac{P}{1-P}\right)=\alpha+\beta x_1+\gamma x_2+ix_1x_2$$

其中, $P$ 表示个体患病概率, $\alpha$ 为群体平均效应, $\beta$、$\gamma$ 和 $i$ 分别为位点 $G$ 和 $H$ 的主效应及它们之间的交互效应。通过似然比检验, 构建近似服从 $\chi^2_{(1)}$分布的检验统计量, 检测交互项系数 $i$ 是否为零, 从而推断两位点的风险等位基因间是否存在交互效应。传统 logistic 回归模型常用于检验两位点间的乘法互作效应。依据两位点间互作模型(加性×加性、显性×显性和隐性×隐性), 遗传效应可分解成如表 1~3 所示。

SPSS、Plink、SAS 等多种统计分析工具可完成基于 logistic 回归模型的基因互作分析。使用者需要

表 1 在加性×加性互作模型下遗传效应的分解

| 位点 $G$ | 位点 $H$ | | |
|---|---|---|---|
| | $H_2/H_2$ | $H_1/H_2$ | $H_1/H_1$ |
| $G_2/G_2$ | $\alpha$ | $\alpha+\gamma$ | $\alpha+2\gamma$ |
| $G_1/G_2$ | $\alpha+\beta$ | $\alpha+\beta+\gamma+i$ | $\alpha+\beta+2\gamma+2i$ |
| $G_1/G_1$ | $\alpha+2\beta$ | $\alpha+2\beta+\gamma+2i$ | $\alpha+\beta+2\gamma+4i$ |

表 2 在显性×显性互作模型下遗传效应的分解

| 位点 $G$ | 位点 $H$ | | |
|---|---|---|---|
| | $H_2/H_2$ | $H_1/H_2$ | $H_1/H_1$ |
| $G_2/G_2$ | $\alpha$ | $\alpha+\gamma$ | $\alpha+\gamma$ |
| $G_1/G_2$ | $\alpha+\beta$ | $\alpha+\beta+\gamma+i$ | $\alpha+\beta+\gamma+i$ |
| $G_1/G_1$ | $\alpha+\beta$ | $\alpha+\beta+\gamma+i$ | $\alpha+\beta+\gamma+i$ |

表 3　在隐性×隐性互作模型下遗传效应的分解

| 位点 $G$ | 位点 $H$ | | |
|---|---|---|---|
| | $H_2/H_2$ | $H_1/H_2$ | $H_1/H_1$ |
| $G_2/G_2$ | $\alpha$ | $\alpha$ | $\alpha+\gamma$ |
| $G_1/G_2$ | $\alpha$ | $\alpha$ | $\alpha+\gamma$ |
| $G_1/G_1$ | $\alpha+\beta$ | $\alpha+\beta$ | $\alpha+\beta+\gamma+i$ |

依互作模型对基因型及模型分量进行正确的编码。由于基于logistic回归模型的基因互作分析方法采用传统的广义线性模型，因此其理论基础好、结果可释性强。Briggs等[4]在类风湿性关节炎的遗传互作研究中，应用logistic回归分析方法成功发现*CDH13*、*MYO3A*、*CEP72* 和*WFDC1* 中的基因变异同*PTPN22* 存在基因互作，并显著影响个体对类风湿性关节炎的易感性。

### 1.2　单体型 **logistic** 回归模型

单体型logistic回归模型实质上是检测等位基因间的交互作用[5]。其基本原理是：非连锁位点间的基因互作会导致位点间等位基因组合(或称单体型)的频率分布在病例和对照间出现差异。与上述基于基因型的模型相似，单体型logistic回归模型可表述如下：

$$\log\left(\frac{P(D=1\mid x_1,x_2)}{1-P(D=1\mid x_1,x_2)}\right)=\alpha+\beta x_1+\gamma x_2+ix_1x_2,$$

式中 $x_1$ 和 $x_2$ 为指示变量，分别表示在单体型构型中两个 SNP 的等位基因信息；$P(D=1|x_1,x_2)$表示携带某单体型 $x_1x_2$ 的个体患病的概率。两位点 4 种单体型对应的模型编码如表 4 所示。通过似然比检验，构建近似服从 $\chi^2_{(1)}$分布的检验统计量，检测交互项系数 $i$ 是否为零，从而推断两位点的风险等位基因间是否存在交互效应。

对于二倍体生物(如人类)，以单体型为分析单位可以增大 1 倍的样本量，因而可提高检测基因互作效应的统计功效。然而应用单体型 logistic 回归时需要注意两点：第一，将基因型数据转化为单体型数据造成个体内两个单体型数据间不独立，违反了 logistic 回归模型的独立性假设；第二，个体单体型的相型信息无法直接通过实验手段观察，需要通过算法(如 EM 算法)进行估计，无可避免地增加互作检验统计量的估计误差。

表 4　单体型变量编码

| 单体型 | 变量 | | |
|---|---|---|---|
| | $x_1$ | $x_2$ | $x_1x_2$ |
| $G_1H_1$ | 1 | 1 | 1 |
| $G_1H_2$ | 1 | 0 | 0 |
| $G_2H_1$ | 0 | 1 | 0 |
| $G_2H_2$ | 0 | 0 | 0 |

## 2　不基于互作模型的统计分析方法

此类方法通过直接构建广义基因互作检验统计量检验互作效应。因为不需要事先假定位点间的互作遗传模型，与基于模型的基因互作分析方法相比，此类方法具有较强的稳健性。

### 2.1　*fast-epistasis* 统计量

*fast-epistasis*统计量是一种快速检测基因互作的方法，由Purcell等[6]于 2007 年推导并整合到著名遗传分析软件Plink(--fast-epistasis命令)中。Plink软件日益受到青睐，其互作检测方法应用广泛。*fast-epistasis*的分析步骤如下：

(1)分别将病例和对照组的两位点基因型联合分布(表 5)整理为相应的 2×2 等位基因列联表(表 6)。

表 5　两位点基因型联合分布

| 位点 $G$ | 位点 $H$ | | |
|---|---|---|---|
| | $H_1H_1$ | $H_1H_2$ | $H_2H_2$ |
| $G_1G_1$ | $n_{22}$ | $n_{21}$ | $n_{20}$ |
| $G_1G_2$ | $n_{12}$ | $n_{11}$ | $n_{10}$ |
| $G_2G_2$ | $n_{02}$ | $n_{01}$ | $n_{00}$ |

表 6　*fast-epistasis* 的等位基因列联表

| 位点 $G$ | 位点 $H$ | |
|---|---|---|
| | $H_1$ | $H_2$ |
| $G_1$ | $a=4n_{22}+2n_{21}+2n_{12}+n_{11}$ | $b=4n_{20}+2n_{21}+2n_{10}+n_{11}$ |
| $G_2$ | $c=4n_{02}+2n_{01}+2n_{12}+n_{11}$ | $d=4n_{00}+2n_{01}+2n_{10}+n_{11}$ |

(2)根据表 6，分别计算病例组和对照组的等位基因对数优势比及其方差，进而构建基因互作检验统计量 $T$：

$$T=\frac{(\hat{\lambda}_A-\hat{\lambda}_N)^2}{\hat{v}_A+\hat{v}_N},$$

其中，$\hat{\lambda}_A$ 和 $\hat{\lambda}_N$ 分别为病例组和对照组的对数优势

比，$\hat{\lambda}_i = \log\frac{ad}{bc}$ ($i$=$A$ 或 $N$)、$\hat{v}_i$ ($i$=$A$ 或 $N$)为 $\hat{\lambda}_i$ 的估计方差。$T$ 服从自由度为 1 的 $\chi^2$ 分布。

*fast-epistasis* 统计量由于不需要估算单体型频率，因此极大降低了计算复杂度，提高了计算效率。但该方法假定组成双杂合基因型组合($G_1G_2/H_1H_2$)的 4 种可能单体型的频率相等，这会低估等位基因间的关联程度；并且在两位点存在连锁不平衡时，会增大犯Ⅰ类错误的概率。

### 2.2 等位基因优势比检验

Wu等[7]提出通过EM算法估算两位点等位基因的联合分布，构建与*fast-epistasis*法相似的统计量 $T_{wu}$：

$$T_{Wu} = \frac{(\hat{\lambda}_A - \hat{\lambda}_N)^2}{\hat{v}_A + \hat{v}_N},$$

其中，参数 $\hat{\lambda}_A$ 和 $\hat{\lambda}_N$ 分别表示病例组和对照组等位基因的对数优势比，$\hat{v}_i$ ($i$=$A$或$N$)为相应的估计方差。$T_{wu}$服从自由度为 1 的$\chi^2$ 分布。考虑到单体型估算会导致方差的低估从而增加检验的假阳性率，Ueki和Cordell[8]在Wu统计量的基础上，提出了 $\hat{v}_i$ 的改进算法，通过一系列仿真实验，表明采用改进的方差估计能较好的控制检测基因互作效应的I类错误率。Wu等[7]通过统计量$T_{wu}$检验方法，在牛皮癣病例对照研究中发现基因*LST1/NCR3*、*CXCR5/ BCL9L*和*GLS2*之间存在互作效应。

### 2.3 *Fst* 方法

Rao等[9]从群体进化角度解释基因互作与疾病之间的关系，提出了基于固定指数(*Fst*)的两位点基因互作检测方法，即将疾病人群和对照人群视为来自相同祖先的不同进化亚群，并采用多元方差分析技术构建基因互作检验统计量*Fst*:

$$Fst=\det(\mathbf{SSW})/\det(\mathbf{SSW}+\mathbf{SSB}),$$

其中，**SSW** 和 **SSB** 分别表示群体内(组内)和群体间(组间)两位点等位基因间的协方差矩阵。在样本量足够大时，*Fst* 渐近服从 Wilks Lambda 分布 $\Lambda(k, n-m, m-1)$，其中 $k$ 是位点数目，$n$ 是单倍体总数，$m$ 为亚群数量。对于基于病例对照研究设计的两位点互作分析($k$=2, $m$=2, $n$ 为样本量的 2 倍)，公式为：

$$F_{Fst} = \frac{(n-m)-1}{m-1}\cdot\frac{1-\sqrt{Fst}}{\sqrt{Fst}} \sim F\left(2, 2(n-3)\right),$$

如果$F_{Fst}>F_{(2,2(n-3))}(\alpha)$，$\alpha$为检验水准，即认为两位点间存在交互作用。本课题组提出方法后，将其应用于疟疾出生队列研究数据，发现*Hb*基因和$\alpha$+-地中海贫血基因间存在互作[9]。

## 3 数据挖掘算法

此类算法认为，互作的 SNP 组合可有效解释疾病的变异程度，具有良好的病例对照分辨能力；算法通过在 SNP 组合搜索空间寻找具有高分类能力的 SNP 组合来识别 SNP 互作。数据挖掘算法多基于机器学习，通常不需要假定遗传互作模型。

### 3.1 随机森林法

随机森林法(Random forest, RF) [10]是一种利用分类树(决策树)对数据进行分类或判别的方法，多适用于结局变量是分类变量的设计类型，如病例对照设计。随机森林是多个决策树的集合，每棵树由自举法(Bootstrap)产生的随机样本训练生成，树的分支由随机选取的属性子集(遗传变异或环境风险因素组合)决定。现有基于Java开发的RF可视化软件RAFT(http://www.stat.berkeley.edu/~breiman/RandomForests/cc_graphics.htm)和在开源统计开发平台R中实现的“ RF ”程序包供研究者选择使用。

对于一个有 $N$ 个个体、$M$ 个 SNP 的数据集，单个决策树的生成步骤如下(图 1)：

(1) 用自举法(Bootstrap)生成样本量为 $N$的随机样本；

(2) 对于树中的每个节点，在 $M$ 个 SNP 中随机选取 $m$($m$<$M$)个 SNP；

(3) 在 $m$ 个 SNP 中选择分割样本效果最佳的 SNP 作为该节点(例如增加了样本子集的同质性)；

(4) 重复第二和第三步，直至终末叶节点出现或满足预先设定的决策树停止生长条件。

如此重复抽样和建模多次，生成多棵决策树，即形成“ 森林 ”。树中的每一个分支都被视为一种SNP组合，组合内的SNP之间存在交互作用和联合作用。随后通过对各种SNP组合预测错误率(Prediction error rate)和重要性得分(Importance score) [11]的计算比较，选择最佳的SNP组合。由于互作的SNP组合可以在更大程度上解释疾病的变异，在病例和对照的分类识别上具有更高的准确性(或更低的预测错误率)，因此该组合倾向于获得更高的重要性得分。

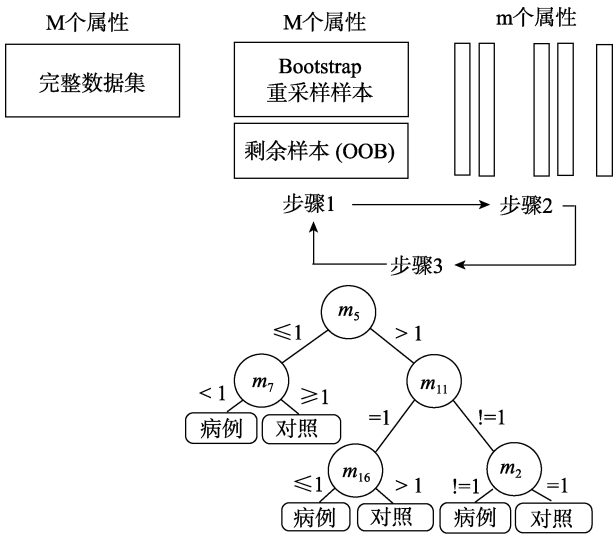


**图 1**　随机森林实现过程

RF法易于发现主效应微弱但与其他SNP有强联合效应的遗传变异[12]，能以较少的参数调整，检验众多预测变量。Lunetta等[13]指出，随机森林法不需要事先定义遗传模型，决策树的早期分割可看作是对群体异质性的剖分和子集提纯的过程，因此可适用于异质性高的复杂疾病的研究。但RF隐含假设认为预测变量之间相互独立。如果高危SNP间存在相关，将会降低这些SNP在RF中的重要性得分，降低了识别功效。此外，由于RF法涉及大量决策树构建及随机抽样过程，因此当样本量庞大及SNP维数很高(如全基因组关联研究)时，RF法会产生巨大的计算负荷，在很大程度上限制了其在高维数据中的推广。

### 3.2　多因子降维法

Ritchie等[14]提出了基于病例对照设计的多因子降维法(Multifactor dimensionality reduction, MDR)用于检测多基因间互作，并提供了免费的MDR应用程序(http://www.multifactordimensionalityreduction.org/)。该方法的基本思想是检索多个位点从低维到高维的全部基因型组合，并判断这些基因型组合的风险高低，来寻找可能影响疾病风险的重要基因型组合。其基本步骤如下：

(1) 将样本随机等分为 $k$ 份，其中 $k$-1 份作为训练集, 1 份为测试集;

(2) 从众多研究因素中选取 $n$ 个 SNP;

(3) 考虑 $n$ 个 SNP 从低维度到高维度(1 维到 $n$ 维)的所有可能基因型组合，划分训练集个体为多个子集;

(4) 计算每个 SNP 基因型组合中的病例/对照频数比，根据预先设定界值(如 1.0)，确定高危 SNP 组合(频数比≥1.0)和低危 SNP 组合(频数比<1.0);

(5) 通过测试集计算每个 SNP 组合的预测错误率(Prediction error rate)和交叉验证一致性(Cross-validation consistency);

(6) 随机置换(Permutation)疾病状态，得到预测错误率的经验分布，进行假设检验，做出统计推断。

但是，Ritchie的MDR方法只能将基因型组合分为高风险和低风险，无法量化其疾病风险。因此有研究者提出改进的MDR法(OR-MDR)[15]。该方法将Ritchie法的第 4 步改进为计算每个联合基因型的风险比，对于$SNP_1$和$SNP_2$，风险比可定义为：

$$OR_{ij}=\frac{P\{SNP_1=i,SNP_2=j\,|\,疾病\}}{P\{SNP_1=i,SNP_2=j\,|\,对照\}}=\frac{P\{疾病\,|\,SNP_1=i,SNP_2=j\}}{P\{对照\,|\,SNP_1=i,SNP_2=j\}}\Big/\frac{P\{疾病\}}{P\{对照\}}$$

其中, $i$ 和 $j$ 分别为 $SNP_1$ 和 $SNP_2$ 的基因型。如果 $OR_{ij}$ 大于 1，则认为该联合基因型是危险因素，反之为保护因素。

MDR属于非参方法，不需要预先设定遗传模型。由于该算法简单易行，对计算平台硬件要求不高，可快速识别高阶基因互作，因此在遗传研究中被广泛使用。Ryan等[16]将其应用在肺结核数据中，在 19 个候选位点中发现rs2305619、rs187084,和rs11465421 的三阶相互作用可影响肺结核的易感性。但该方法在样本量较少时会增加犯假阳性或假阴性错误的概率。此外，高阶SNP组合容易出现对照组频数为 0 的情况，此时MDR无法计算风险比，因此MDR方法一般适用于阶数不太高(如≤5 阶)的基因互作识别。

### 3.3　遗传规划法

遗传规划算法(Genetic programming, GP) [17]是在遗传算法(Genetic algorithm, GA)的基础上发展起来的智能寻找问题最优解的一种进化算法。其基本思想是，随机产生一个适合于给定环境的初始群体;群体中的每个个体均为问题的一个解决方案，都有一个对于给定问题的适应度值；个体经过选择、变

异和交换产生下一代，保留高适应度个体，如此进化下去，直到获得给定问题的最优解或近似解。基于Matlab 的工具箱 GPLAB(http://gplab.sourceforge.net 可下载)可实现 GP 算法。GP 可用于 SNP 高阶互作效应分析，具体步骤如下：

(1) 随机产生初始群体，群体中每个个体均为随机产生的高阶 SNP 组合，这里的 SNP 即真实数据中的 SNP 变量。例如，图 2 描述的个体 $L$=(SNP$_1$=3)∨((SNP$_2$≠1)∧(SNP$_3$=1))(注：1~3 为 SNP 基因型编码，分别代表野生型、杂合型和突变纯合型；∨和∧分别表示 OR(或)和 AND(并)逻辑关系)，其中，SNP$_1$=3、SNP$_2$≠1 和 SNP$_3$=1 为 3 个表达式。如果 $L$ 值为真，即 SNP$_1$=3 和(SNP$_2$≠1)∧(SNP$_3$=1)中至少有一个为真，此时观察对象划为病例，否则为对照；

(2) 计算每个个体的适应度，淘汰适应度低的个体；

(3) 根据预设概率参数执行交叉和突变运算，产生新群体，计算新群体中各个体的适应度，择优保留；

(4) 重复步骤 3，直至满足终止条件。选择适应度最优的个体作为遗传规划的结果；

(5) 对第 4 步中的结果评价其错误分类率和表达式简洁性。权衡后选出满足条件的SNP基因型组合。

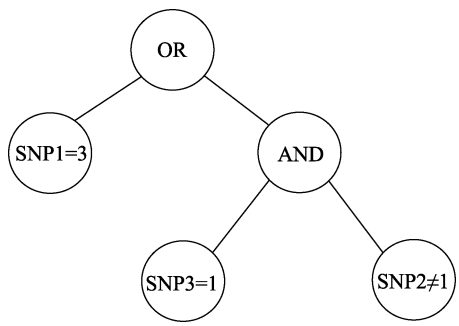


图 2 个体 L 的逻辑表达式

GP算法根据生物界自然选择和进化机制，通过模拟生物个体中的基因行为，在SNP互作空间中搜索高分辨能力的SNP基因型组合。相对于遗传算法，在个体表示方法上，GP算法克服了传统遗传算法中固定长度二进制字符串表示形式的局限，使用了更为灵活的表现方式。GP算法得到的最优个体可看作是高阶互作的SNP组合，这对临床上进行疾病诊断和大规模筛查技术的开展具有指导价值。Liu等[18]将传统GP进行改进，对芯片数据成功实现了特征选择和分类。

近年来，也见一些其他方法应用于SNP/基因互作分析，包括基于遗传算法的优化神经网络方法(GPNN)[19]、支持向量机(Support vector machine)[20]等；这些方法识别基因互作的基本思想与上述方法有相似之处，读者可参考相关文献。

## 4 基因互作分析方法间的比较

早期的应用于分类性状的基因互作分析方法建立在传统的统计模型分析的基础上，适用于简单的低阶基因互作分析，因而有其局限性(表 7)。这类方法，如 logistic 回归，随着互作阶数的增加，会面临

表 7 几种基因互作分析方法的特性和优缺点

| 方法 | 模型依赖 | 高阶互作检测 | 优点 | 缺点 | 实现工具 | 实例 |
|---|---|---|---|---|---|---|
| LRM | √ | × | 结果易解释 | 不适用高维 SNP 互作 | SPSS; PLINK | 类风湿关节炎[4] |
| Fast-epistasis | × | × | 不需估算单倍型; 计算复杂度低 | 无法剖分互作效应和联合效应; 主效应和 LD 会增加假阳性率 | PLINK | 牛皮癣[7] |
| Fst | × | × | 计算复杂度低; 功效较高 | 无法剖分互作效应和联合效应 | R | 疟疾[9] |
| RF | × | √ | 可分析稀疏数据; 发现非线性互作 | 受主效应影响; SNP 间相关性会影响检验效能; 无法剖分互作效应和联合效应 | RAFT; R | 哮喘病[11] |
| MDR | × | √ | 有效降维; 发现非线性互作 | 无法剖分互作效应和联合效应; 搜索高通量 SNP 数据的效率低; 不适用于小样本 | MDR | 肺结核[16] |
| GP | × | √ | 可分析稀疏数据; 发现非线性互作; | 过拟合; 计算复杂度高 | GPLAB | 乳腺癌[17] |

注： “√”表示“依赖”、“适用”； “×”表示“不依赖”、“不适用”。

“维度灾难”的困扰。不基于互作模型的统计方法多数是通过构建两个位点间基因互作或联合效应的测度实现对两位点间互作的快速检测，但尚未推广到三维或以上的高维基因互作分析。

基于机器学习的生物信息学算法一般无需事先假定遗传模型，适用于分析稀疏的高维数据和SNP间非线性互作关系[21]，但也有其不足。数据挖掘类算法通过搜索变量组合以寻求病例对照的有效分类，并没有清晰剖分SNP互作效应和联合效应。基于树模型的RF法在构造单棵树时依赖于边际作用显著的SNP，会导致主效应和低阶互作效应微弱情况下树结构的不稳定性。MDR算法思想较为简单，可快速搜索所有可能的基因型组合，但其仅能定性或简单量化特定SNP基因型组合的风险，尚缺乏对SNP组合互作模式或性质的合理量化。GP算法基于进化理论，模拟SNP在人类群体中的进化历程，并从中搜索解释疾病风险的最佳SNP组合。其关键在于有效缩小算法搜索空间，加快其收敛速度。

综上所述，应用于高通量遗传学数据的基因互作分析方法众多，各存优劣[21, 22]。单一方法难以解决全基因组数据的各种复杂情况。综合利用各种方法，建立集成分析系统，有助于更全面的挖掘复杂疾病的基因互作效应。

## 5　结语与展望

基因间复杂的交互作用在人类常见复杂疾病的发生发展中发挥重要作用。尽管基因互作分析方法已经有了一定的发展，但目前全基因组范围内互作检测仍面临不少的挑战和一些亟待解决的问题。

### 5.1　连续型性状中的基因互作

近年来也有人关注连续型性状基因互作分析方法。由于连续型性状的数值特性，其传统的基因互作分析中，多采用多重线性回归，通过对回归模型中互作项回归项系数的假设检验，来检测互作效应的统计显著性[23]；类似分类性状，也有人尝试通过构造U检验统计量达到快速检测连续型性状基因互作效应的目的[24]。另外，可简单地将连续型性状转化为分类性状，采用前述的几类互作检测方法进行分析，但需要指出的是这种做法会造成数据信息的损失，降低互作效应的检验功效。

### 5.2　统计学互作的生物学解释

基因互作常被称为上位互作。Bateson[25]提出生物学互作的定义，认为上位效应是两位点共同影响同一性状时，其中一个位点的等位基因能够抑制另一个位点的等位基因的表现；Fisher[26]从统计学角度提出上位效应是两位点效应中偏离其独立可加效应的部分。Moore和Williams[27]指出，统计互作和生物互作定义的不同需引起研究者注意：生物互作是发生在分子水平的可导致性状改变的现象，而统计互作是基于人群遗传变异数据的基因型-表型的统计关联。统计学互作并不必然揭示生物学互作，但统计学互作往往为生物学互作的发现提示方向。统计互作的解释离不开生物学知识和证据的支持，在建立统计互作与生物学互作的联系时，可借助功能学数据库先验知识，增加其生物学解释的能力。

### 5.3　互作模型的不确定性

基于模型的互作效应检测关键在于位点遗传效应的合理剖分。但是，根据研究假设的不同，互作模型可有多种定义，例如根据是否显现单基因效应可分为无边际效应模型和有边际效应模型；根据外显率模型不同又可分为上位模型、异质性模型、阈值模型等[3]。Hallgrimsdottir等[28]则从几何角度定义了69 种两位点互作模型。互作模型的不确定性会极大影响互作检测的功效。此外，现有的互作模型定义存在向高阶推广的难度。Wade指出$n$个基因中$k$维互作的正交回归项的数目为$\binom{n}{k}\times 2^k$，这在全模型的拟合中难以实现[14]。高阶互作模型除了要考虑模型参数个数，还要考虑零假设模型的构建，用以检验不同维度的互作效应的存在。

### 5.4　互作效应和联合效应的混淆

Wan等[28]指出，允许互作效应存在的关联分析，可以发现主效应微弱但互作效应显著的SNP，但是避免不了主效应显著引发联合效应显著的情况。以两位点为例，它们的联合效应可以看成是位点的主效应和位点间交互效应的一个函数。主效应或互作效应的出现和增加均可导致联合效应的上升。值得注意的是，目前多数模型非依赖的方法，特别是基于机器学习的互作识别算法(如MDR)实质上检测的

是两位点的联合效应而非纯粹的互作效应。两种效应的混淆会导致由主效应引起的假阳性率上升，影响结果的真实性。精细剖分联合效应中的互作效应是目前机器学习算法的一个重要课题。

### 5.5 全基因组互作检测的计算难题

Cordell[1]报道，在单节点计算机中使用Plink软件，对 89 294 个SNP的所有两两互作效应进行检测需要花费 14 天时间。即使在 3GHz处理器和 4G内存下，对 5 000 例样本和 10 000 个SNP的两两互作检测亦需要 3 个多小时[28]。在应对三阶或更高阶互作时，时间开销会更大。高维SNP互作分析对算法的效率和计算机的性能提出了更高的要求。发展高效快速的高维互作分析算法势在必行，高性能并行计算程序的开发可大大提高互作检测的效率[29]。此外，发展具有广泛适应性的稳健方法，可以加快基因间复杂的非线性交互作用的检测。McKinney等[30]认为，基于网络的方法可以有效整合互作效应和主效应的信息，可更深层次发掘基因互作与疾病易感的关系和影响因素。例如，Hu等[31, 32]提出基于网络的模型优化算法，成功构造了膀胱癌的互作基因网络，并对该网络进行三维互作分析，提示了基于网络的基因互作研究的潜力。

### 5.6 SNP/基因互作的下游功能学分析

SNP/基因互作的下游功能学分析尚处于严重滞后状态。生物学先验知识的快速积累和各种分子生物学数据库的日趋完善，为全面和深入挖掘SNP/基因互作的生物学机制，识别对疾病易感性有鉴别力的分子特征组合提供了在生物功能层次上的解析。常用的数据库包括SNP数据库(dbSNP、HapMap计划等)[33]、基因注释系统(GENE ONTOLOGY, http://www.geneontology.org/)、通路数据库(KEGG)[34]、疾病数据库(OMIM)[35]等。充分利用以上信息，开展SNP-基因、SNP-代谢通路、SNP-功能模块的下游功能学分析，是实现从构建SNP/基因互作分析到互作基因和分子靶点挖掘再到功能网络分析的全程式系统生物学分析过程的关键所在。

## 参考文献(References):


[1] Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*, 2009, 10(6): 392–404.[DOI]

[2] Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered*, 2003, 56(1–3): 73–82.[DOI]

[3] Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, 2005, 37(4): 413–417. [DOI]

[4] Briggs FBS, Ramsay PP, Madden E, Norris JM, Holers VM, Mikuls TR, Sokka T, Seldin MF, Gregersen PK, Criswell LA, Barcellos LF. Supervised machine learning and logistic regression identifies novel epistatic risk factors with PTPN22 for rheumatoid arthritis. *Genes Immun*, 2010, 11(3): 199–208. [DOI]

[5] Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet*, 2003, 73(6): 1316–1329. [DOI]

[6] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, 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]

[7] Wu XS, Dong H, Luo L, Zhu Y, Peng G, Reveille JD, Xiong MM. A novel statistic for genome-wide interaction analysis. *PLoS Genet*, 2010, 6(9): e1001131. [DOI]

[8] Ueki M, Cordell HJ. Improved statistics for genome-wide interaction analysis. *PLoS Genet*, 2012, 8(4): e1002625. [DOI]

[9] Rao SQ, Yuan MQ, Zuo XY, Su WY, Zhang F, Huang K, Lin MH, Ding YL. A novel evolution-based method for detecting gene-gene interactions. *PLoS ONE*, 2011, 6(10): e26435. [DOI]

[10] Breiman L. Random forests. *Machine Learning*, 2001, 45(1): 5–32. [DOI]

[11] Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P. Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol*, 2005, 28(2): 171–182. [DOI]

[12] Cook N, Zee R, Ridker P. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat Med*, 2004, 23(9): 1439–1453. [DOI]

[13] Lunetta K, Hayward L, Segal J, Van Eerdewegh P. Screening large-scale association study data exploiting interactions using random forests. *BMC Genet*, 2004, 5: 32. [DOI]

[14] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD,

Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 2001, 69(1): 138–147. [DOI]

[15] Chung YJ, Lee SY, Elston RC, Park T. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. *Bioinformatics*, 2007, 23(1): 71–76. [DOI]

[16] Collins RL, Hu T, Wejse C, Sirugo G, Williams SM, Moore JH. Multifactor dimensionality reduction reveals a three-locus epistatic interaction associated with susceptibility to pulmonary tuberculosis. *BioData Min*, 2013, 6(1): 4. [DOI]

[17] Nunkesser R, Bernholt T, Schwender H, Ickstadt K, Wegener I. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics*, 2007, 23(24): 3280–3288. [DOI]

[18] Liu KH, Xu CG. A genetic programming-based approach to the classification of multiclass microarray datasets. *Bioinformatics*, 2009, 25(3): 331–337. [DOI]

[19] Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioinformatics*, 2003, 4(1): 28. [DOI]

[20] Chen SH, Sun JL, Dimitrov L, Turner AR, Adams TS, Meyers DA, Chang BL, Zheng SL, Gronberg H, Xu JF, Hsu FC. A support vector machine approach for detecting gene-gene interaction. *Genet Epidemiol*, 2008, 32(2): 152–167. [DOI]

[21] McKinney BA, Reif D, Ritchie M, Moore JH. Machine learning for detecting gene-gene interactions. *Appl Bioinformatics*, 2006, 5(2): 77–88. [DOI]

[22] Chen L, Yu GQ, Langefeld CD, Miller DJ, Guy RT, Raghuram J, Yuan XG, Herrington DM, Wang Y. Comparative analysis of methods for detecting interacting loci. *BMC Genomics*, 2011, 12(1): 344. [DOI]

[23] Bocianowski J. A comparison of two methods to estimate additive-by-additive interaction of QTL effects by a simulation study. *J Theor Biol*, 2012, 308: 20–24. [DOI]

[24] Li M, Ye C, Fu W, Elston RC, Lu Q. Detecting genetic interactions for quantitative traits with U-statistics. *Genet Epidemiol*, 2011, 35(6): 457–468. [DOI]

[25] Bateson W, Mendel G. Mendel's Principles of Heredity. Cambridge: Cambridge University Press, 1909. [DOI]

[26] Fisher RA. The correlation between relatives on the supposition of mendelian inheritance. *Trans Roy Soc Edinb*, 1919, 52(2): 399–433. [DOI]

[27] Moore JH, Williams SM. Epistasis and its implications for personal genetics. *Am J Hum Genet*, 2009, 85(3): 309–320. [DOI]

[28] Hallgrímsdottir IB, Yuster DS. A complete classification of epistatic two-locus models. *BMC Genet*, 2008, 9: 17. [DOI]

[29] 李放歌, 王志鹏, 户国, 李辉. 全基因组关联研究中的交互作用研究现状. 遗传, 2011, 33(9): 901–910. [DOI]

[30] McKinney BA, Pajewski NM. Six degrees of epistasis: statistical network models for GWAS. *Front Genet*, 2011, 2: 109. [DOI]

[31] Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, Moore JH. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*, 2011, 12(1): 364. [DOI]

[32] Hu T, Andrew A, Karagas M, Moore J. Statistical epistasis networks reduce the computational complexity of searching three-locus genetic models. *Pac Symp Biocomput*, 2013: 397–408. [DOI]

[33] Glinsky GV. Integration of HapMap-based SNP pattern analysis and gene expression profiling reveals common SNP profiles for cancer therapy outcome predictor genes. *Cell Cycle*, 2006, 5(22): 2613–2625. [DOI]

[34] Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. *Nucleic Acids Res*, 2002, 30(1): 42–46. [DOI]

[35] McKusick VA. Mendelian inheritance in man and its online version, OMIM. *Am J Hum Genet*, 2007, 80(4): 588–604. [DOI]