遗传 Hereditas (Beijing) 2018 年 2 月, 40(2): 162－169
www.chinagene.cn

稀有变异遗传关联性研究中常用负担检验方法比较

林欣琪[1,2]，梁融[1]，张俊国[1]，皮路程[1]，陈思东[1]，刘丽[1]，郜艳晖[1]

1. 广东药科大学公共卫生学院流行病与卫生统计学系，广州 510310
2. 广东省职业病防治院，广州 510300

摘要：为比较稀有变异遗传关联研究中常用负担检验方法(CMC、WST、SUM 及其扩展)在不同遗传情境下的统计性能，本文通过计算机模拟产生不同样本量、连锁不平衡(linkage disequilibrium, LD)参数、混杂非关联变异的个数和不同效应的关联变异等条件的稀有变异病例对照数据集，运用各种负担检验方法进行分析，分别计算各方法的一类错误和效能。结果表明，各方法一类错误均在 0.05 附近；当稀有变异效应方向一致时，除 aSUM 法外，LD 参数越大、混杂非关联变异越少、各法效能越高；当效应方向不一致时，各法效能则显著降低。除强 LD 外，有方向考虑的方法效能均比无方向考虑的方法高，且样本量越大效能越高。负担检验的统计性能受效应大小和方向、噪音变异和连锁不平衡等多种因素影响。在实际应用中，在各类方法选择、确定集合单位，权重等时最好结合遗传变异的生物信息先验以提高研究效能。

关键词：稀有变异；遗传关联研究；负担检验

Comparison of common burden tests for genetic association studies of rare variants

Xinqi Lin[1,2], Rong Liang[1], Junguo Zhang[1], Lucheng Pi[1], Sidong Chen[1], Li Liu[1], Yanhui Gao[1]

1. Department of Epidemiology and Biostatistics, School of Public Health, Guangdong Pharmaceutical University, Guangzhou 510310, China
2. Guangdong Province Hospital for Occupational Disease Prevention and Treatment, Guangzhou 510300, China

Abstract: Common burden tests have different statistical performance in genetic association studies of rare variants. Here, we compare the statistical performance of burden tests, such as CMC, WST, SUM and extension methods, using the computer-simulated datasets of rare variants with different parameters of sample sizes, linkage disequilibrium (LD), and different numbers of mixed non-associated variants. The simulation results showed that the type I error for all methods is near 0.05. When the rare variants had the same direction of effect, the higher LD and the less non-associated variants, the

收稿日期: 2017−11−12; **修回日期:** 2017−12−28
基金项目: 广东省自然基金(编号：2016A030313809)，广东省科技厅公益能力(编号：2014A020212307)和国家自然科学青年基金(编号：81302493)资助[Supported by the National Natural Science Foundation of Guangdong, China (No. 2016A030313809), Science and Technology Planning Project of Guangdong Province, China (No. 2014A020212307), and National Natural Science Foundation of China (No.81302493)]
作者简介: 林欣琪，硕士研究生，研究方向：分子流行病学。E-mail: linxinki@163.com
通讯作者: 刘丽，博士，副教授，研究方向：流行病与卫生统计学。E-mail: pupuliu919@163.com
DOI: 10.16288/j.yczz.17-174
网络出版时间: 2018/1/18 11:23:00
URI: http://kns.cnki.net/kcms/detail/11.1913.R.20180118.1112.002.html

higher the power of these method, except the data adaptive SUM test. When the direction was different, the power was significantly reduced for all methods. The methods that consider the direction yielded larger statistical power than those methods without considering the effect direction, except the strong LD condition. And the larger the sample size, the larger the power. The statistical performance of burden tests is affected by a variety of factors, including the sample size, effect direction of variants, non-associated variants, and LD. Therefore, when choosing the method and setting the collection unit and weight, the prior biological information of genetic variation should be integrated to improve study efficiency.

Keywords: rare variant; genetic association studies; burden test

随着二代测序技术的迅猛发展，稀有变异(rare variant)对复杂性状的作用越来越受到关注[1, 2]。由于稀有变异在人群中次等位基因频率(minor allele frequency, MAF)很低(<5%)，传统单变量或多变量检验方法效能极低。为克服此问题，研究者常把感兴趣区域(region of interest, ROI)内稀有变异集合成遗传负担得分直接用于关联分析，统称为负担检验(burden test)。如对病例和对照集合后的新变量进行检验的 CAST 法(cohort allelic sums test)[3]，或将稀有变异集合后的遗传得分与常见变异一起进行 *Hotelling* T^2 检验的 CMC 法(combined multivariate and collapsing)[4]，采用变异频率的方差进行加权的 WST 法(weighted sum test)[5]，或将协变量与集合后的稀有变异同时纳入回归模型的 SUM 法(sum test)，以及在 SUM 基础上考虑变异效应方向的 SSU 法(sum of the squares of the marginal score statistics)，SSUw 法(weighted form of sum of the squares of the marginal score statistics)[6,7]和 aSUM 法(data adaptive sum test)[8]。但各类方法的统计性能及应用效果仍有待于进一步研究。为此，本文通过计算机模拟病例-对照遗传关联研究数据，比较常用负担检验方法在各种遗传情境时的效能和一类错误，为负担检验在稀有变异遗传关联分析中的有效应用提供参考和依据。

1　原理与方法

目前较常用的负担检验包括 CMC、WST、SUM 及其扩展(SSU, SSUw 和 aSUM)。

1.1　CMC 法

CMC 法根据等位基因频率或不同的 ROI 把遗传区域分成 k 个亚区域，并在每个亚区域内集合所有稀有变异位点，再进行 *Hotelling's* T^2 检验，统计量为

$$T^2 = \frac{n_1 n_0}{n_1 + n_0}(\bar{X}^A - \bar{X}^{\bar{A}})^T S^{-1}(\bar{X}^A - \bar{X}^{\bar{A}}) \quad (1)$$

式(1)中 n_1 和 n_0 分别为病例组和对照组样本量，$\bar{X}^A$ 和 $\bar{X}^{\bar{A}}$ 分别为两组集合后的稀有变异均数向量，S 为方差协方差阵。在稀有变异与疾病无关的假设下，T^2 近似服从非中心、自由度为 k 的 χ^2 分布。

1.2　WST 法

假设频率较低的变异的遗传效应可能更大，Madsen 和 Browning[5]提出先根据总样本量 n 和对照组的变异频率 q_j 计算 ROI 内第 j 个变异的权重 $\hat{w}_j = \sqrt{n \cdot q_j(1-q_j)}$，再计算个体 i 的遗传得分，

$$\gamma_i = \sum_{j=1}^{p} \frac{X_{ij}}{\hat{w}_j} \quad (2)$$

式(2)中 X_{ij} 表示个体 i 在变异位点 j 上的等位基因突变数，如加性遗传模型时 $X_{ij} \in \{0,1,2\}$。类似于 Wilcoxon 检验，将所有个体的遗传得分 γ_i 排序，得到病例和对照组遗传得分的秩和，并采用置换检验得到 p 值。

1.3　SUM 法及扩展

Pan[9]假设 ROI 内所有稀有变异与疾病的关联强度相等，回归系数为 β_c，则称为 SUM 法，

$$Logit\,\Pr(Y_i = 1) = \beta_{c0} + \sum_{j=1}^{p} X_{ij}\beta_c \quad (3)$$

并采用得分检验(score test)进行统计推断。

$$T_{Score} = U'V^{-1}U \tag{4}$$

式(4)中得分向量及协方差分别为

$$U = \sum_{i=1}^{n}(Y_i - \bar{Y})X_i,$$

$$V = \bar{Y}(1-\bar{Y})\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})' \tag{5}$$

SUM 法仅估计一个参数 β_c，大大减少自由度。但当稀有变异作用方向不一致时，导致效能降低。为此将 SUM 法改进以克服效应方向不同的影响, 如根据边际得分统计量的平方和构造新统计量, 称为 SSU,

$$\mathrm{SumSqU} = U_M'U_M = (Y-\bar{Y})'XX'(Y-\bar{Y}) \tag{6}$$

式(6)中 U_M 为边际得分统计量。进一步考虑其方差协方差，可采用加权形式，即 SSUw,

$$\mathrm{SumSqUw} = \mathrm{U}_{\mathrm{M}}'\mathrm{Diag}(V)^{-1}U_M \tag{7}$$

式(6)和(7)检验统计量的无效分布是一个二项式，可由 $a\chi_d^2 + b$ 近似估计得到。可以看到，二次项对遗传变异的效应方向并不敏感。

此外，数据自适应(data adaptive)的方式也自然地被提出[8]，在给定检验水准 α_0 下先根据单因素回归结果对保护效应的遗传变异进行反向编码，如 $\hat{\beta}_{M,j} < 0$ 且 $P_{M,j} \leqslant \alpha_0$，则 $X_{\cdot j}$ 变成 $X_{\cdot j}^* = 2 - X_{\cdot j}$，否则 $X_{\cdot j}^* = X_{\cdot j}$。对重新编码后的数据拟合模型(3)，计算得分统计量 U 及其方差 V，最后通过置换样本构造 aSum 的检验统计量,

$$\alpha Sum = (U-U_0)'V_0^{-1}(U-U_0) \tag{8}$$

式(8)中 U_0 和 V_0 分别为 B 个置换样本得分统计量和方差的均数；小样本时，根据无效分布 $a\chi_b^2 +$ 计算 p 值，其中 a 和 b 根据 Satterthwaite 近似估计[10]。

2 模拟实验

2.1 模拟数据的参数设置

本文模拟病例对照研究数据，假设病例数 n_1 等于对照数 n_0，设稀有变异个数为 8，指定 OR(odds ratio)均为 1，即稀有变异与疾病状态均无关联时用于计算一类错误；指定 OR 均为 2 时表示关联变异的效应方向相同；4 个 OR 为 2、其余 4 个为 0.5 时表示关联变异的效应方向不同。设置 $OR_j = \exp b_j$，其中 $b_j = \dfrac{1}{\sqrt{n \cdot MAF_j(1-MAF_j)}}$ 时，表示对 MAF 越小的变异设置越大的效应。其它参数还有样本量(n_1=n_0=250、500 和 1000)、变异间连锁不平衡(linkage disequilibrium, LD)参数(ρ=0、0.5 和 0.9)、混杂非关联变异(OR 固定为 1)的个数(0、4、8 和 16)。疾病的基线患病率定义为 $p_0 = 0.05$。每种模拟条件下共产生 1000 个数据集。

2.2 模拟数据集的生成

在上述模拟实验条件下产生病例对照模拟数据集，具体步骤[9, 11]为：

(1)产生服从多元正态分布的向量 $Z = (Z_1, \cdots, Z_p)'$，定义任两元素间的关联 $Corr(Z_i, Z_j) = \rho^{|i-j|}$；(2)将向量 Z 转化为二分类的单倍体，关联变异和非关联变异单倍体的 MAF 分别从均匀分布 $U(0.001, 0.01)$ 和 $U(0.01, 0.05)$ 中随机抽样，根据 MAF 值计算对应正态分布下的分位数，记为界值。如果 $\boldsymbol{Z}$ 中对应元素小于该界值，定义单倍体 X1_1 中对应元素为 1，否则为 0；(3)循环步骤(2)产生另一单倍体 X1_2。X1_1 与 X1_2 相加得到每个稀有变异的基因型数据(0, 1, 2)；(4)个体 i 的疾病状态 Y_i 根据 logistic 回归模型产生,

$$p(Y_i = 1) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j\right)} \tag{9}$$

式(9)中，$\beta_j = \ln(OR_j)$，$\beta_0 = \ln\left(\dfrac{p_0}{1-p_0}\right)$，$OR_j$ 和 p_0 如前述定义。

2.3 统计分析

对每种条件下的 1000 个模拟数据集，均采用 CMC、WST、SUM 及其扩展方法(SSU, SSUw 和 aSUM)进行分析，除 WST 外，对 ROI 内稀有变异集合的方式采用计数赋值法(ROI 内稀有变异突变的总数作为新变量)。分别计算各种模拟条件下各法的一类错误或效能。数据模拟及统计分析均采用 R V3.0.2 软件，调用 Assoteste R 包进行 SUM 及扩展方法，CMC 和 WST 应用 Basu 和 Pan[7]的 R 程序。

3　结果与分析

3.1　各类方法的一类错误

表 1 可见，无论样本量如何改变，各种连锁不平衡情况和不同非关联稀有变异数量对各种方法的一类错误无太大影响, 均在 0.05 水平附近 ;除 aSUM 法在非关联稀有变异数量为 16 时, 一类错误值异常小幅度增大, 最高可达 0.071, 其他情形中各方法一类错误均保持在 0.05 水平上下。总的来说，随着样本量的增加，大多数方法的一类错误变化不大，均在 0.05 水平附近 ;随着非关联稀有变异数量的改变, aSUM 法的一类错误存在小幅度的增大。

3.2　各类方法的效能

稀有变异效应方向一致时，随着连锁不平衡参数从 0 到 0.9，各方法的效能逐渐增大，$\rho = 0.9$ 时大部分方法效能达到 1，而 aSUM 法在 $\rho = 0.9$，样本量增加至 1000 时，效能为 0.009~0.022，存在明显的降幅；除 $\rho = 0.9$ 外，随着非关联稀有变异数量增多，各方法的效能逐渐下降，但其对效能值的影响小于连锁不平衡参数。各种方法效能值的具体比较发现，除 aSUM 法外，LD 参数越大，各方法效能越高；混杂非关联变异越多，各方法效能越低。小样本、无连锁不平衡且混杂非关联变异数较少时，无方向考虑的方法(CMC、WST 和 SUM)和 aSUM 的效能值在无非关联变异或数量较小时效能值较其他方法高；而大样本且连锁不平衡程度高时，aSUM 效能极低(表 2)。

稀有变异效应方向不一致时，各法效能显著降低。随着连锁不平衡参数从 0 到 0.9，各种方法的效能均逐渐降低；随着非关联稀有变异数量增多，各

表 1　各模拟条件下各类方法的一类错误

Table 1　Type I error in various methods under the different simulations

$n_1=n_0$	方法	非关联变异数=0			非关联变异数=4			非关联变异数=8			非关联变异数=16		
		$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
250	CMC	0.055	0.046	0.055	0.030	0.063	0.036	0.045	0.051	0.042	0.046	0.036	0.047
	WST	0.048	0.057	0.054	0.032	0.046	0.033	0.058	0.047	0.045	0.039	0.048	0.044
	SUM	0.049	0.040	0.037	0.031	0.056	0.037	0.055	0.050	0.041	0.038	0.050	0.051
	SSU	0.054	0.045	0.050	0.033	0.053	0.044	0.052	0.044	0.048	0.053	0.054	0.051
	SSUw	0.057	0.045	0.060	0.033	0.051	0.032	0.055	0.042	0.053	0.057	0.047	0.048
	aSUM	0.052	0.058	0.046	0.040	0.058	0.058	0.063	0.057	0.047	0.055	0.053	0.067
500	CMC	0.048	0.048	0.053	0.035	0.054	0.034	0.053	0.045	0.042	0.043	0.051	0.052
	WST	0.046	0.048	0.055	0.039	0.056	0.039	0.047	0.047	0.048	0.033	0.051	0.058
	SUM	0.040	0.047	0.054	0.039	0.051	0.047	0.050	0.046	0.048	0.036	0.041	0.050
	SSU	0.058	0.047	0.054	0.047	0.075	0.046	0.058	0.052	0.047	0.059	0.049	0.050
	SSUw	0.047	0.051	0.049	0.039	0.073	0.052	0.052	0.053	0.051	0.056	0.042	0.042
	aSUM	0.061	0.055	0.063	0.052	0.053	0.045	0.060	0.048	0.051	0.048	0.046	0.071
1000	CMC	0.043	0.048	0.059	0.052	0.048	0.040	0.044	0.041	0.047	0.036	0.055	0.048
	WST	0.045	0.052	0.052	0.052	0.050	0.041	0.038	0.050	0.051	0.043	0.053	0.053
	SUM	0.044	0.052	0.060	0.046	0.054	0.050	0.047	0.039	0.049	0.041	0.057	0.055
	SSU	0.056	0.048	0.061	0.049	0.054	0.047	0.045	0.053	0.045	0.040	0.048	0.056
	SSUw	0.044	0.045	0.060	0.052	0.053	0.049	0.041	0.056	0.047	0.051	0.051	0.054
	aSUM	0.047	0.048	0.064	0.054	0.051	0.045	0.050	0.051	0.056	0.047	0.058	0.070

ρ表示连锁不平衡参数；稀有变异效应值 OR 均为 1。

表 2 关联稀有变异效应一致时各类方法的效能
Table 2 Results of power in various methods under the same effects with associated rare variants

$n_1=n_0$	方法	非关联变异数=0			非关联变异数=4			非关联变异数=8			非关联变异数=16		
		$\rho=0$	$\rho=0.5$	$\rho=0.9$	$\rho=0$	$\rho=0.5$	$\rho=0.9$	$\rho=0$	$\rho=0.5$	$\rho=0.9$	$\rho=0$	$\rho=0.5$	$\rho=0.9$
250	CMC	0.734	0.813	0.999	0.564	0.747	0.996	0.349	0.687	0.970	0.316	0.618	0.906
	WST	0.725	0.900	1.000	0.577	0.856	0.999	0.446	0.771	0.998	0.362	0.732	0.997
	SUM	0.756	0.962	1.000	0.607	0.943	1.000	0.511	0.914	1.000	0.398	0.889	1.000
	SSU	0.457	0.914	1.000	0.386	0.917	1.000	0.345	0.906	1.000	0.335	0.906	1.000
	SSUw	0.421	0.918	1.000	0.339	0.902	1.000	0.286	0.895	1.000	0.271	0.884	1.000
	aSUM	0.592	0.903	1.000	0.427	0.875	0.998	0.340	0.829	1.000	0.242	0.814	0.999
500	CMC	0.946	0.991	1.000	0.858	0.974	1.000	0.660	0.960	1.000	0.573	0.940	1.000
	WST	0.947	0.992	1.000	0.859	0.977	1.000	0.739	0.965	1.000	0.621	0.933	1.000
	SUM	0.950	0.998	1.000	0.883	0.998	1.000	0.764	0.993	1.000	0.667	0.992	1.000
	SSU	0.764	0.998	1.000	0.734	0.995	1.000	0.679	0.991	1.000	0.645	0.993	1.000
	SSUw	0.733	0.998	1.000	0.678	0.994	1.000	0.603	0.985	1.000	0.511	0.991	1.000
	aSUM	0.899	0.997	0.725	0.787	0.991	0.748	0.655	0.982	0.755	0.542	0.983	0.728
1000	CMC	1.000	0.999	1.000	0.984	1.000	1.000	0.942	1.000	1.000	0.892	0.995	1.000
	WST	1.000	1.000	1.000	0.987	0.999	1.000	0.961	0.998	1.000	0.881	0.995	1.000
	SUM	1.000	1.000	1.000	0.991	1.000	1.000	0.963	1.000	1.000	0.908	1.000	1.000
	SSU	0.973	1.000	1.000	0.968	1.000	1.000	0.941	1.000	1.000	0.921	1.000	1.000
	SSUw	0.969	1.000	1.000	0.963	1.000	1.000	0.925	1.000	1.000	0.888	1.000	1.000
	aSUM	0.995	0.921	0.011	0.979	0.951	0.009	0.945	0.967	0.022	0.856	0.985	0.011

ρ表示连锁不平衡参数；关联稀有变异效应值 OR 均为 2。

种方法的效能也逐渐降低；具体来看，LD 程度和非关联变异个数对各法效能影响不大；除高度连锁不平衡时外，有方向考虑的方法(SSU、SSUw 和 aSUM)效能值均比无方向考虑的方法(CMC、WST 和 SUM)高，且样本量越大效能越高(表 3)。

稀有变异效应与 MAF 有关(MAF 越小，效应越大)时，随着连锁不平衡参数变大，各方法的效能逐渐升高；当非关联稀有变异数和连锁不平衡参数均为 0 时，无方向考虑的简单折叠法(CMC、WST 和 SUM)效能比有方向考虑的方法(SSU、SSUw 和 aSUM)高，达 0.5 以上；当没有非关联变异时，WST 效能略高于 CMC，但随着非关联变异数量增多，前者效能略低于后者。整体上各法效能比效应一致时低；且强连锁不平衡及非关联变异数量少时效能较高，但和样本量无明显关系；不存在非关联变异时，WST 方法的效能并不低于 CMC 和 SUM 方法，但存在非关联变异时，随着非关联变异数量增多，WST 效能降低(表 4)。

4 讨 论

负担检验在稀有变异遗传关联研究中因其原理朴素、思路简单已有较多应用，但其统计性能受效应大小和方向、噪音变异和连锁不平衡等多种因素影响。已有研究[7]对多种稀有变异统计检验(包括负担检验、基于自适应聚类检验、C-α检验和基于重复检验)进行模拟试验，在无 LD 或有 LD、样本量为 500、不同非关联变异个数和效应的不同组合情况下，表明负担检验和基于自适应聚类检验的效能最大。本文中将不同的 LD 参数，样本量和非关联变异个数交叉组合成 36 种实验条件，分别比较了每种条件下各法的 I 类错误及在效应方向一致、不一致、效应与 MAF 有关时的效能，结果显示在本研究模拟条件下各法一类错误均可接受，但在不同遗传情境下效能有很大的变化。在相同的部分模拟条件下，本文结果与文献[7]类似。

最初提出的负担检验 CMC 和 SUM 假设集合内

表 3　关联稀有变异效应方向不同时各类方法的效能

Table 3　Results of power in various methods under the different effects with associated rare variants

n_1=n_0	方法	非关联变异数=0			非关联变异数=4			非关联变异数=8			非关联变异数=16		
		ρ = 0	ρ = 0.5	ρ = 0.9	ρ = 0	ρ = 0.5	ρ = 0.9	ρ = 0	ρ = 0.5	ρ = 0.9	ρ = 0	ρ = 0.5	ρ = 0.9
250	CMC	0.118	0.199	0.139	0.090	0.184	0.103	0.171	0.133	0.082	0.162	0.151	0.075
	WST	0.120	0.124	0.145	0.092	0.131	0.135	0.074	0.091	0.111	0.071	0.089	0.099
	SUM	0.123	0.129	0.123	0.099	0.142	0.112	0.076	0.098	0.088	0.072	0.110	0.117
	SSU	0.367	0.331	0.167	0.298	0.306	0.150	0.268	0.227	0.113	0.230	0.244	0.134
	SSUw	0.329	0.285	0.134	0.255	0.255	0.091	0.240	0.190	0.095	0.187	0.156	0.118
	aSUM	0.276	0.261	0.145	0.225	0.222	0.144	0.163	0.164	0.111	0.136	0.159	0.126
500	CMC	0.170	0.340	0.248	0.144	0.323	0.214	0.323	0.279	0.165	0.291	0.282	0.183
	WST	0.169	0.210	0.233	0.137	0.194	0.199	0.081	0.141	0.176	0.088	0.125	0.196
	SUM	0.172	0.218	0.170	0.156	0.189	0.181	0.095	0.142	0.161	0.095	0.168	0.193
	SSU	0.669	0.608	0.247	0.595	0.539	0.241	0.518	0.475	0.190	0.480	0.423	0.228
	SSUw	0.638	0.565	0.207	0.562	0.487	0.222	0.482	0.427	0.163	0.414	0.337	0.159
	aSUM	0.610	0.529	0.258	0.485	0.444	0.260	0.396	0.351	0.200	0.318	0.294	0.230
1000	CMC	0.283	0.595	0.403	0.249	0.553	0.390	0.524	0.525	0.335	0.525	0.517	0.341
	WST	0.271	0.338	0.331	0.251	0.288	0.309	0.167	0.228	0.278	0.139	0.211	0.280
	SUM	0.293	0.342	0.266	0.253	0.293	0.267	0.172	0.243	0.243	0.143	0.258	0.270
	SSU	0.941	0.909	0.410	0.894	0.845	0.373	0.856	0.785	0.323	0.828	0.736	0.326
	SSUw	0.921	0.893	0.348	0.884	0.829	0.315	0.828	0.750	0.286	0.776	0.658	0.282
	aSUM	0.891	0.802	0.436	0.775	0.696	0.407	0.682	0.618	0.342	0.575	0.504	0.346

ρ表示连锁不平衡参数；关联稀有变异效应值中 4 个 OR 为 2，4 个为 0.5。

表 4　MAF 越小稀有变异效应越大时各类方法的效能

Table 4　Results of power in various methods under the less MAF while more effects with associated rare variants

n_1=n_0	方法	非关联变异数=0			非关联变异数=4			非关联变异数=8			非关联变异数=16		
		ρ = 0	ρ = 0.5	ρ = 0.9	ρ = 0	ρ = 0.5	ρ = 0.9	ρ = 0	ρ = 0.5	ρ = 0.9	ρ = 0	ρ = 0.5	ρ = 0.9
250	CMC	0.596	0.695	0.998	0.215	0.548	0.979	0.217	0.415	0.949	0.189	0.314	0.905
	WST	0.605	0.803	0.999	0.276	0.508	0.942	0.181	0.367	0.898	0.156	0.296	0.837
	SUM	0.620	0.914	1.000	0.239	0.817	1.000	0.154	0.729	0.999	0.112	0.714	0.999
	SSU	0.289	0.828	1.000	0.085	0.658	1.000	0.075	0.591	0.996	0.066	0.610	0.999
	SSUw	0.297	0.850	1.000	0.225	0.841	1.000	0.185	0.778	1.000	0.165	0.761	1.000
	aSUM	0.487	0.858	1.000	0.232	0.769	0.999	0.151	0.667	0.999	0.105	0.639	0.999
500	CMC	0.540	0.683	1.000	0.190	0.553	0.995	0.230	0.414	0.988	0.169	0.343	0.964
	WST	0.545	0.766	0.999	0.239	0.427	0.916	0.199	0.310	0.874	0.157	0.250	0.799
	SUM	0.560	0.890	1.000	0.211	0.764	1.000	0.158	0.675	1.000	0.107	0.637	0.998
	SSU	0.272	0.783	1.000	0.056	0.550	1.000	0.072	0.519	1.000	0.054	0.539	0.997
	SSUw	0.272	0.813	1.000	0.212	0.803	1.000	0.195	0.738	1.000	0.150	0.707	1.000
	aSUM	0.455	0.831	1.000	0.211	0.692	1.000	0.146	0.596	1.000	0.109	0.581	0.999
1000	CMC	0.506	0.619	0.997	0.181	0.482	0.991	0.212	0.303	0.980	0.183	0.374	0.946
	WST	0.511	0.700	0.998	0.232	0.259	0.841	0.171	0.287	0.767	0.146	0.227	0.699
	SUM	0.542	0.837	1.000	0.200	0.675	0.999	0.144	0.645	0.996	0.096	0.585	0.998
	SSU	0.230	0.724	1.000	0.065	0.470	0.999	0.057	0.430	0.995	0.069	0.467	0.996
	SSUw	0.259	0.767	1.000	0.201	0.727	1.000	0.182	0.719	1.000	0.161	0.654	0.999
	aSUM	0.428	0.778	1.000	0.190	0.611	1.000	0.127	0.563	0.997	0.109	0.508	0.997

ρ表示连锁不平衡参数；关联稀有变异效应值 $OR_j = \exp(b_j) = \exp\left(\frac{1}{\sqrt{n \cdot MAF_j(1-MAF_j)}}\right)$。

稀有变异效应大小和方向一致，WST 假设变异的 MAF 越低效应越强，实际研究中这些假设时常违背。本研究显示当集合内稀有变异效应方向一致时，除强 LD 大样本时的 aSUM 法外，其他情况时各方法均有较高效能，但把具有相反效应的稀有变异集合在一起，效能则大大降低。aSUM 方法本适用于效应方向不一致的情况，在效应方向一致且强 LD 的大样本时，和其他负担检验相比，aSUM 需先进行单变异检验，因此有可能对由于抽样误差导致的部分变异进行反向编码，从而抵消了集合变量的效应导致效能降低。当我们增加了模拟样本量为 2000 时强 LD 的情况，在非关联变异个数为 0、4、8 和 16 时其效能分别是 0.0145、0.012、0.022 和 0.0135，这和样本量为 1000 时的结果类似。当效应大小和 MAF 有关时，随着非关联变异增多，WST 法效能降低，可能的原因是模拟数据集中，非关联变异也从一定 MAF 范围内随机抽取，在 WST 法分析时同时根据非关联变异的 MAF 赋予了相应的权重，因此可能增加了非关联变异的效应，而影响了集合后关联变异的效能。而后续发展的 SSU、SSUw 和 aSUM 等方法在大样本且低 LD 时效能虽有改善，但多数情况仍效果不佳。因此，在实际稀有变异关联分析时，如何确定稀有变异分析集合是提高统计效能最为关键的问题之一。Nicolae[12]建议根据各种信息来源确定集合，包括(1)根据基因物理位置，如某基因的所有外显子区域变异；(2)基于变异功能(如移码，错义或终止)只集合有害的外显子变异亚集；(3)利用生物信息学工具计算出功能得分较大的外显子变异亚集；(4)综合多种资源信息的得分获得亚集。此外，本文只考虑了变异效应方向一致和不一致的情况，而理论和模拟研究[8]均表明效应方向一致时效应越大各法效能越高；而方向不一致时，有方向考虑的方法效能可能更稳健，但更全面的结论需要更为系统的模拟研究。

此外，变异间的 LD 对负担检验的效能也有很大影响。和弱 LD 相比，强 LD 的变异效应方向一致时负担检验的集合策略增强了遗传效应；但效应方向相反时，各法效能显著降低。LD 导致病例间有更高的 DNA 序列相似性，因此，除负担检验外，另一类基于个体间遗传相似性的方差分量检验方法如 C-α[13]，SKAT(sequence kernel association test)[14, 15]等，将集合内稀有变异的作用看作随机效应，将检验病例和对照组间变异频率的差别转化为检验随机效应的方差。此类方法理论上不受集合中稀有变异方向不同的影响，对混杂较多非关联变异也较为稳健。实际应用时，可利用文献、生物信息数据库等来源的先验信息确定稀有遗传变异间的 LD、效应方向(保护效应或危险效应)和 MAF 值等条件再加以选择统计方法。如有证据表明研究的变异中同时有保护效应和危险效应时，需要考虑不依赖效应方向的 SSU 和 SSUw/aSUM 等方法；如同为保护性效应或危险效应但存在强连锁不平衡时，可选择除 aSUM 外的其他负担检验；如变异功能注释未发现有力的功能学证据时，不应将这些变异再纳入关联研究或统计分析，以保证方法的效能。此外，本研究报告的各种遗传情景下不同样本量时的效能表也可为实际工作中稀有变异关联研究的样本量和效能估计提供参考。

参考文献(References)：

[1] Lettre G. Rare and low-frequency variants in human common diseases and other complex traits. *J Med Genet*, 2014, 51(11): 705–714. [DOI]

[2] Wu L, Schaid DJ, Sicotte H, Wieben ED, Li H, Petersen GM. Case-only exome sequencing and complex disease susceptibility gene discovery: study design considerations. *J Med Genet*, 2014, 52(1): 10–16. [DOI]

[3] Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Muta Res/Fundam Mol Mechan Mutag*, 2007, 615(1-2): 28–56. [DOI]

[4] Li BS, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*, 2008, 83(3): 311–321. [DOI]

[5] Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS*

Genet, 2009, 5(2): e1000384. [DOI]

[6] Pan W, Shen XT. Adaptive tests for association analysis of rare variants. *Genet Epidemiol*, 2011, 35(5): 381–388. [DOI]

[7] Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol*, 2011, 35(7): 606–619. [DOI]

[8] Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered*, 2010, 70(1): 42–54. [DOI]

[9] Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol*, 2009, 33(6): 497–507. [DOI]

[10] Satterthwaite F. An approximate distribution of estimates of variance components. *Biomet Bull*, 1946, 2(6): 110–114. [DOI]

[11] 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]

[12] Nicolae DL. Association tests for rare variants. *Annu Rev Genomics Hum Genet*, 2016, 17(7): 117–130. [DOI]

[13] Neale B, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. *PLoS Genet*, 2011, 7(3): e1001322. [DOI]

[14] Zhang T. An introduction to support vector machines: and other kernel-based learning methods. *AI Magazine*, 2001, 22(2): 103–104. [DOI]

[15] Wu MC, Lee S, Cai TX, Li Y, Boehnke M, Lin XH. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*, 2011, 89(1): 82–93. [DOI]

(责任编委: 方向东)