·基础理论与方法·

# 低频变异关联研究与统计检验

廖鑫 邓静 荀佳雨 严俊霞
410078 长沙，中南大学湘雅公共卫生学院流行病与卫生统计学系
通信作者：严俊霞，Email：20457456@qq.com
DOI：10.3760/cma.j.issn.0254-6450.2017.07.026

【摘要】 在过去的10年间，基于“常见疾病-常见变异”的假设，全基因组关联研究被广泛应用于疾病和复杂性状的遗传学病因研究中。但是，全基因组关联分析发现的疾病相关常见变异，只能解释疾病小部分的遗传风险，造成“遗传度丢失”。“常见疾病-低频变异”的假设被提出。随着新一代测序技术的发展，低频变异关联研究陆续开展。本文主要对低频变异关联研究的研究设计以及关联分析方法进行综述。

【关键词】 低频变异；关联研究；统计检验
**基金项目：**国家自然科学基金（81502881）；中国博士后科学基金（2015M582351）

**Less common variants association study and statistical analysis** *Liao Xin, Deng Jing, Xun Jiayu, Yan Junxia*
*Department of Epidemiology and Health Statistics, Xiangya School of Public Health, Central South University, Changsha 410078, China*
*Corresponding author: Yan Junxia, Email: 20457456@qq.com*

【**Abstract**】 In the past decade, based on the “common disease-common variant” hypothesis, genomewide association studies (GWAS) have been extensively used to dissect the genetic components of complex diseases and quantitative traits. However, the identified disease-associated common variants by GWASs can only explain small part of the corresponding disease heritability. “Common disease-rare variant” hypothesis has been proposed to explore the missed heritability. With the development of the next generation sequencing technology, various association studies of less common variants are ongoing. This paper provides an overview of the study designs and statistical tests of these variants.

【**Key words**】 Less common variants; Association study; Statistical test
**Fund programs:** National Natural Science Foundation of China (81502881); China Postdoctoral Science Foundation (2015M582351)

过去10年间基于“常见疾病-常见变异”的假说，全基因组关联分析（genome wide association study，GWAS）被广泛应用于疾病和复杂性状的遗传学病因研究中[1]。以芯片为基础，利用病例对照研究设计，基于连锁不平衡原理，在全基因组范围内对常见变异［最小等位基因频率（Minor allele frequency，MAF）＞0.05］与疾病之间的关联进行探讨。传统GWAS研究不涵盖低频变异（MAF＜0.05）（不存在或者只存在微弱连锁不平衡），造成疾病“遗传度丢失”（missing heritability）[2]。迄今为止，GWAS研究发现的众多疾病相关常见变异，具有较小的致病效应（*OR*s：1.1～1.5），仅能解释疾病小部分的遗传风险[3]。

有研究从自然进化的角度阐释了人类遗传变异产生的过程，认为常见变异出现较早并经受住了选择压力，大部分倾向于中性效应；但低频变异出现较晚，尚未在人类进化过程中被淘汰，更有可能是功能性的，在疾病发生发展中起重要作用[4]。近年来，随着二代测序技术的发展，多项研究发现基因低频变异与糖尿病、痛风、炎性肠炎、心肌梗死等复杂疾病相关，且效应较高（*OR*s＞2）[5-8]，支持“常见疾病-低频变异”假说[9]。目前，低频变异与疾病和健康关系的研究正陆续开展。本文主要对低频变异关联研究的研究设计以及关联分析方法进行综述。

## 一、低频变异关联研究

遗传变异可位于基因组的任何区域。理想情况下，可利用传统的流行病学病例对照研究或者队列研究设计，进行全基因组深度测序，探讨低频变异与

疾病的关联。然而，全基因组深度测序费用昂贵，大样本研究难以开展。目前常采用成本相对较低的全基因组低深度测序、靶向测序、外显子测序、外显子芯片等进行低频变异关联研究。

1. 全基因组低深度测序：测序深度是指测序得到的碱基总量与基因组大小的比值，测序深度低，测序错误率较高。但是，模拟实验表明，在测序成本一定的情况下，相对于小样本的深度测序，大样本低深度测序无论是在变异检测还是疾病关联分析中都具有更强的效能[10]。Li 等[11]研究表明，对于MAF＞0.002的变异，在单变异关联检验中，在4×低深度测序3 000个样本与30×深度测序2 000个样本的效能是相同的。如利用全基因组低深度测序，Gudmundsson 等[12]对53名冰岛前列腺癌患者和1 742名对照进行测序发现了新的前列腺癌易感位点；Converge 等[13]对5 303名中国女性重度抑郁症患者和5 337名对照测序发现了两个抑郁症易感位点。全基因组低深度测序节约了成本，但样本量较小时会使测序错误率升高，注意样本量的选择。目前，基于测序的低频变异关联研究效能和样本量的计算方法可见参考文献[14]。

2. 外显子组测序：利用序列捕获技术将外显子区域DNA捕获、富集后进行高通量测序分析。测序结果通过与参考数据库比对、过滤，并通过生物学功能预测、关联分析等，最终确定低频变异易感位点。外显子组约占人类基因组序列的1%～2%，但包含了约85%的致病突变[15]。外显子测序有较高的测序深度，成本较低，能经济有效地发现易感变异。近几年，利用外显子测序先后发现了Miller综合征、Kabuki综合征、脊髓小脑变性等单基因疾病的致病变异[16-18]，并发现了早发性心肌梗死、迟发型阿尔兹海默病等复杂疾病易感低频变异[8,19]。目前，利用外显子测序，一些大型遗传流行病学研究项目正在进行，如美国为研究心脏、肺、血液等相关疾病的遗传学病因而进行的6 500人的外显子测序项目[20-21]；英国为研究低频变异与疾病和健康的关系，对患有各种疾病和症状的6 000例及4 000例健康个体进行测序的万人外显子项目[22]。但是，外显子测序只能发现编码区域的遗传变异，对一些非编码区的变异以及一些结构性变异无能为力。

3. 目标区域靶向测序：是基于DNA杂交原理，将目标基因区域DNA进行富集、测序。可对连锁分析锁定的目标范围或经过全基因组筛选的特定基因或区域进行测序，快速、全面地测出目标区域内相关突变。如Johansen 等[23]对438例高脂血症患者和327名对照进行了GWAS位点靶向测序，发现大量高脂血症相关低频易感变异；Rivas 等[24]对350例克罗恩患者和350名对照的56个候选基因进行了测序，发现了与克罗恩病有关的一些低频和罕见变异。但是，目标区域靶向测序需要事先确定测序区域，不能发现非目标区域的遗传变异。

4. 客户定制的基因分型芯片：是基于GWAS和测序研究的高优先级变异开发的芯片，例如针对代谢和心血管疾病的Metabochip芯片和针对自身免疫和炎症性疾病的Immunochip芯片[25-26]，这些芯片包含一些被筛选出来的可能与疾病相关的常见和低频变异，如功能性的非同义突变、剪切突变、终止子突变，以及之前GWAS发现的SNP、线粒体SNP、人类白细胞抗原标签SNP等，可以验证最初的GWAS发现，以及对GWAS提示的区域进行深入研究。外显子芯片成本较低，能够进行大样本研究。Permuth 等[27]利用外显子芯片对8 165例卵巢上皮癌（EOC）患者和11 619名对照进行基因分型，发现*ACTBL2*、*BTD*等罕见变异与EOC相关。但由于外显子芯片大都是基于欧洲人的数据开发的，在非欧洲人群中代表性差，且芯片只涵盖已知SNPs，不能发现人群特异性的新的低频和罕见变异[28]。

5. 极端表型抽样测序：是对最有可能携带信息的个体进行测序以提高研究效能。常有两种研究设计：基于家族性患者及基于极端表型患者的测序分析。家族性患者富集遗传学信息，是宝贵的遗传学资源，但是发病率低，难以收集。对极端表型个体进行测序更为常用。在疾病研究中，可以已知的危险因素选择具有极端表型的个体，如在2型糖尿病的遗传学病因研究中，病例可选择发病年龄早、低BMI、和（或）有2型糖尿病家族史的个体，而对照可选择年龄大、肥胖、无糖耐量受损的个体进行测序。在数量性状研究中，可选择校正了已知协变量的影响后，具有极端特征值的个体，如表型分布的上下10%。如Emond 等[29]为研究囊性纤维化患者中慢性铜绿假单胞菌感染相关易感变异，利用极端表型抽样分别对7岁前患有慢性铜绿假单胞菌感染的囊性纤维化患者和14岁后仍未感染慢性铜绿假单胞菌的囊性纤维化患者进行外显子测序，发现*CAV2*和*TMC6*基因遗传变异与囊性纤维化患者慢性铜绿假单胞菌感染相关。但此类研究对极端值敏感、存在抽样偏倚、潜在性状需要满足正态性分布等，研究结果不能直接推广到所有人群。

二、低频变异关联分析的统计学方法

低频变异关联研究面临巨大挑战。首先，在人群中观察到一个高概率的低频变异需要大样本量；其次，除非样本量巨大或致病效应极强，传统的以单个变异为基础的关联分析统计效率低下[20]。为提高统计学检验效能，针对低频变异的统计学方法相继被提出。其中以评估基因或区域上多个变异累积效应的聚合检验（aggregation tests）最为常用，它包括负荷检验、自适应负荷检验、方差成分检验、结合负荷检验和方差成分检验的混合检验等。相应的统计软件包及下载地址可参考相关文献[30]。

1. 负荷检验（burden test）：负荷检验假定基因或者特定区域内的所有低频变异与某一性状存在因果关联，并且方向相同，效应大小一致（调整权重后）。在此前提下，将多个遗传变异的信息转化成一个遗传评分，然后对评分和性状进行关联分析。计算遗传评分的一个简单方法是对基因或者特定区域内所有变异的最小等位基因计数。遗传评分可根据不同的发病机制假设进行定义，如在队列等位基因加和检验（cohort allelic sums test，CAST）中，假设任何低频变异都增加疾病风险，如果给定区域内没有最小等位基因，记为0，反之为1[31]。另外，考虑到不同频率变异的遗传效应不同，可根据变异的MAF对遗传评分进行加权。此外，还可根据变异的生物学信息对变异评分进行加权[32]。目前，Polyphen2、SIFT等在线工具可以预测变异的生物学影响，如可将变异分为“良性或可耐受变异”“可能有害变异”“有害变异”等。关联分析时，可以仅对“可能有害变异”和/或“有害变异”进行分析，也可以对不同生物学影响的变异赋予不同的权重进行分析[33]，然后利用回归模型计算相关风险及检验统计量和*P*值。Lange等[34]对564位LDL-C高于98百分位水平的个体，391位LDL-C低于2百分位水平的个体以及3 307名对照进行外显子测序，利用负荷检验发现*PNPLA5*基因上的低频变异与LDL-C水平相关。

2. 自适应负荷检验（adaptive burden tests）：不同变异可能有不同的作用方向（致病作用、无关联或保护作用），简单的负荷分析未考虑这一因素，出现这种情况时统计检验效能将会大幅度降低。为提高检验效能解，研究者提出了几种改良方法，如基于核函数的自适应聚类加权检验（kernel-based adaptive cluster，KBAC）、估计回归系数检验（estimated regression coefficient test，EREC）、数据自适应加和检验（data adaptive sum test，aSum）、set-up检验等。KBAC利用基于核函数的自适应加权将风险和非风险变异分类及关联分析结合起来。如研究者利用这种方法对达拉斯心脏研究数据进行分析[35]，新发现了一些与能量代谢相关的罕见变异[36]；EREC检验直接估计每个变异的回归系数并作为权重，由于当最小等位基因数较小时，$\beta_i$估计值不稳定，其检验统计量只在大样本量的情况下准确，所以运用Bootstrap估计*P*值。研究者利用EREC对Colaus心血管研究数据进行分析[37]，发现了与TC相关低频变异[38]；aSum检验，首先对低频变异进行单变量logistic回归，求回归系数和*P*值，对*P*值小于事先定义的检验水准且作用方向相反的变异反向编码赋值，之后进行负荷检验，采用置换检验估计*P*值[39]；Step-up检验在模型选择框架下先筛选掉可能无关联的变异再定义权重[40]。aSum和Step-up在真实数据模拟试验中都展现了良好的自适应能力及检验效能。此外，其他数据自适应方法还有可变阈值检验（variable threshold，VT）[33]，以及具有自适应加权方案的加权加和方法（weighted sum test，WST）等[41]。

自适应负荷检验对每个变异的潜在遗传结构进行的假设更少，比原始负荷检验更强大。但它对个体变异进行回归系数估计非常困难，并且对低频变异不稳定。且大多数自适应检验需要用置换检验估计*P*值，计算量很大[30]。

3. 方差成分检验（variance-component tests）：负荷检验及改良的自适应负荷检验重在比较不同表型间（病例组*vs.*对照组）低频变异频率上的差别，但如果与表型相关的低频变异并非表现为频率上的差异，而主要表现为DNA序列上的不同，即出现连锁不平衡的情况，负荷检验统计效能将大为降低。为此，研究者们开发出了以检验变异频率分布的方差为基础的方差成分检验，如：C-alpha检验[42]、序列核关联性检验（sequence kernel association test，SKAT）[43-44]以及方差得分合计检验（sum of squared score test，SSU）等[45]。C-alpha检验通过比较携带遗传变异者中病例频率的实际方差与二项分布理论模型下的期望方差，检验特定基因或者区域内混有大量无关联变异时是否存在关联变异，它能识别相同频率均数下方差的变化，当目标检验区域内有大量无关变异时有更高的检验效能。如Faino等[46]对GAW18高血压数据进行C-alpha检验发现了与高血压相关的低频变异。此外，它不受遗传变异作用方向和效应值的影响，比负荷检验稳健，但是缺点是不能调整协变量，且仅适用于定量性状。为克服C-alpha检验的缺

陷，研究者提出SKAT检验，在回归模型中引入代表遗传变异效应的核函数项，该核函数测量了任意2个个体间目标区域内遗传变异序列的遗传相似性，采用混合效应模型框架下的方差分量计分检验，检验统计量是单变量得分统计量的加权平方和。SKAT检验近似服从混合$\chi^2$分布，其$P$值可以快速计算分析，可应用于定量性状和二元性状[47]。日本的一项家族性颅内动脉瘤研究对150例家族性颅内动脉瘤患者和150例年龄、性别匹配的非颅内动脉瘤对照进行测序，利用SKAT检验确定了*PKD1*和*PKD2*基因上的罕见非同义突变与颅内动脉瘤相关[48]。对于二元性状，当样本量较小或最小等位基因数较小时，基于大样本量的方法都会产生不准确的Ⅰ型错误率。要在SKAT中计算准确的$P$值，需要应用重采样方法，如置换检验，大大增加了计算量。为此，最近Hasegawa等[49]开发了一种自适应SKAT（adjusted sequence kernel association test，AP-SKAT），在样本量较小的情况下，保证检验效能的同时有效控制Ⅰ型错误率，并且在较短的时间内获得与置换检验一致的$P$值估计，目前该方法尚未应用于实际，但模拟试验表明这种方法具有与SKAT，优化SKAT（Optimal-SKAT，SKAT-O）相当的效能且计算时间更短。在低频变异与环境存在交互作用时，采用负荷检验会产生分析偏倚，为此，Lin等[50]提出了iSKAT（interaction sequencekernel association test），利用加权回归校正协变量，很好的控制了低频变异的主效应。研究者利用iSKAT对CoLaus研究数据进行分析，发现*ADIPOQ*基因上存在血浆脂联素相关低频变异并且与饮酒存在交互作用[51]。阮培峰[52]提出一种适应家系数据的SKAT模型（adjusted-SKAT，ADSKAT），通过对SKAT的原模型进行修改，加入表示家系结构的随机作用向量，使得家系数据中亲属相关性的影响被考虑进模型，并且得出新的检验统计量对应的概率分布。研究者对GAW18高血压数据集进行模拟试验表明，在家系结构的数据中ADSKAT检验效能高于SKAT。

4. 结合负荷检验和方差成分检验的混合检验（combined burden and variance-component tests）：如果一个分析区域内存在较多非关联变异或变异作用方向不一致，方差成分检验要比负荷检验效能更强；反之，若存在较多作用方向一致的关联变异，负荷检验要更强大。但是，实际工作中，往往缺少变异效应的先验信息，在关联分析方法选择时也可将负荷检验和方差成分检验结合起来。目前已提出一些结合负荷检验和方差成分检验的混合检验方法，如Derkach等[53-54]提出的用Fisher法将2种检验的$P$值结合，并用置换法评价检验的显著性，Fisher统计量为$-2\log(P_{SKAT})-2\log(P_{burden})$，其中$P_{SKAT}$和$P_{burden}$分别是SKAT和burden检验的$P$值。Lee等[55-56]提出SKAT-O将SKAT与负荷检验统计量进行线性组合，运用计算效率高的一维数值积分计算SKAT-O的近似$P$值。如Cruchaga等[19]对2 363例迟发型阿尔茨海默患者和2 024名对照进行外显子测序，利用SKAT-O检验，发现*PLD3*基因上的一系列低频有害变异。若真实情况与负荷检验或方差成分检验的假设条件相符，进行混合检验会导致效能降低。但由于遗传变异的先验信息不能确定，所以结合两种检验方法的混合检验是一种可尝试的选择。

5. 连锁混合检验（combined association in the presence of linkage test，CAPL）：负荷检验和SKAT最初是为病例对照设计开发的，之后应用到家系设计中。连锁混合检验（CAPL）提出全新的统计量$T_i$，在处理家系数据时考虑了亲属相关性的影响，再将统计量$T_i$与负荷检验结合（CAPL-Burden）或SKAT结合（CAPL-SKAT），应用到病例对照研究中，利用Boostrap方法估计$P$值。将CAPL-Burden和CAPL-SKAT混合应用，可以减少检验次数提高检验效能。如研究者对GAW19高血压家系和病例–对照结合数据利用CAPL检验，发现了10个高血压易感基因[57]。随着测序研究的深入，这种检验方法对于从家系和病例–对照结合数据中发现候选基因很有帮助。

6. 效能计分合计检验（sum of powered score test，SPU）：当有大量非关联低频变异存在时，这些无关变异会增加检验次数，许多目前已提出的统计方法检验效能会大幅降低。为此，Pan等[58]提出了一种新的检验方法：效能计分合计检验。这种方法基于广义回归的得分向量，可以处理不同类型的性状以及校正协变量。SPU的检验效能取决于未知的待检验低频变异的关联模式所决定的参数$\gamma$值。Pan等[58]在提出SPU时，也提出了自适应SPU，从候选$\gamma$值中选择$P$值最小的进行检验。如研究者从GAW17数据中选择2 476个基因（至少含有1个低频变异）上的18 131个低频变异进行检验，SPU和aSPU有很高的检验效能且当有大量非关联低频变异时，aSPU的自适应能力要强于KBAC，EREC，SKAT-O[58]。SPU，aSPU都是基于高斯似然统计的检验方法，对于有极端值的数据处理效能不强。Wei

等[59]开发了一种Huber损失函数框架下的aSPU检验(robust aSPU test, aSPUr),以解决离群值出现的问题。

7. 指数组合检验(exponential combination test, EC test):指数组合检验是在贝叶斯框架下形成的,检验统计量基于单变量得分统计量平方的指数和,而且零假设的分布未知,所以*P*值的估计需要用置换检验。因为指数函数的增长速度很快,所以只有在非常小部分的变异是因果变异的情况下,指数组合检验的检验效能强于负荷和方差成分检验,否则,检验效能将会降低。Chen等[60]将EC检验应用于铂化合物(广泛应用的抗癌药物)的药物基因组学研究中,揭示了其潜在的耐药性和毒性机制。

另外,还有基于惩罚模型的方法、传统的多变量主成分分析方法、基于模型的多因子降维法等。尽管各类方法在研究效能、Ⅰ类错误率、适用范围及稳健性等方面还存在很多问题,但随着生物信息学的发展及后续实验研究的推进,对低频(罕见)变异的认识也将逐步深入,相应的统计学方法也日益完善和发展。随着外显子组和全基因组大数据库的出现[22, 61],大大提高了低频(罕见)变异关联研究的精确度和可信度,越来越多的低频(罕见)致病变异将被发现,期待对复杂性状或疾病的遗传学病因理解越来越全面和清晰。

**利益冲突** 无

## 参 考 文 献


[1] Visscher P, Brown M, McCarthy M, et al. Five years of GWAS discovery[J]. Am J Hum Genet, 2012, 90(1): 7–24. DOI: 10.1016/j.ajhg.2011.11.029.

[2] Maher B. The case of the missing heritability[J]. Nature, 2008, 456(7218): 18–21. DOI: 10.1038/456018a.

[3] Macarthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog)[J]. Nucleic Acids Res, 2016, 45(D1): D896–901. DOI: 10.1093/nar/gkw1133.

[4] Raychaudhuri S. Mapping rare and common causal alleles for complex human diseases[J]. Cell, 2011, 147(1): 57–69. DOI: 10.1016/j.cell.2011.09.011.

[5] Nejentsev S, Walker N, Riches D, et al. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes[J]. Science, 2009, 324(5925): 387–389. DOI: 10.1126/science.1167728.

[6] Sulem P, Gudbjartsson DF, Walters GB, et al. Identification of low-frequency variants associated with gout and serum uric acid levels[J]. Nat Genet, 2011, 43(11): 1127–1130. DOI: 10.1038/ng.972.

[7] Steinthorsdottir V, Thorleifsson G, Sulem P, et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes[J]. Nat Genet, 2014, 46(3): 294. DOI: 10.1038/ng.2882.

[8] Do R, Stitziel NO, Won H, et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction[J]. Nature, 2014, 518(7537): 102–106. DOI: 10.1038/nature13917.

[9] Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing[J]. Nat Rev Genet, 2010, 11(6): 415–425. DOI: 10.1038/nrg2779.

[10] Morrison AC, Voorman A, Johnson AD, et al. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol[J]. Nat Genet, 2013, 45(8): 899–901. DOI: 10.1038/ng.2671.

[11] Li Y, Sidore C, Kang HM, et al. Low-coverage sequencing: implications for design of complex trait association studies[J]. Genome Res, 2011, 21(6): 940–951. DOI: 10.1101/gr.117259.110.

[12] Gudmundsson J, Sulem P, Gudbjartsson DF, et al. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer[J]. Nat Genet, 2012, 44(12): 1326–1329. DOI: 10.1038/ng.2437.

[13] Converge Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder[J]. Nature, 2015, 523(7562): 588–591. DOI: 10.1038/nature14659.

[14] Wang GT, Li B, Santos-Cortez RPL, et al. Power analysis and sample size estimation for sequence-based association studies[J]. Bioinformatics, 2014, 30(16): 2377–2378. DOI: 10.1093/bioinformatics/btu296.

[15] Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery.[J]. Nat Rev Genet, 2011, 12(11): 745–755. DOI: 10.1038/nrg3031.

[16] Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder[J]. Nat Genet, 2010, 42(1): 30–35. DOI: 10.1038/ng.499.

[17] Ng SB, Bigham AW, Buckingham KJ, et al. Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome[J]. Nat Genet, 2010, 42(9): 790–793. DOI: 10.1038/ng.646.

[18] Shadrina MI, Shulskaya MV, Klyushnikov SA, et al. *ITPR1* gene p. Val1553Met mutation in Russian family with mild Spinocerebellar ataxia[J]. Cerebellum Ataxias, 2016, 3: 2. DOI: 10.1186/s40673-016-0040-8.

[19] Cruchaga C, Karch CM, Jin SC, et al. Rare coding variants in the Phospholipase D3 gene confer risk for Alzheimer's disease[J]. Nature, 2014, 505(7484): 550–554. DOI: 10.1038/nature12825.

[20] Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes[J]. Science, 2012, 337(6090): 64–69. DOI: 10.1126/science.1219240.

[21] Fu W, O'Connor TD, Jun G, et al. Analysis of 6 515 exomes reveals the recent origin of most human protein-coding variants[J]. Nature, 2013, 493(7431): 216–220. DOI: 10.1038/nature11690.

[22] The UK10K Consortium. The UK10K project identifies rare variants in health and disease[J]. Nature, 2015, 526(7571): 82–90. DOI: 10.1038/nature14962.

[23] Johansen CT, Wang J, Lanktree MB, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia[J]. Nat Genet, 2010, 42(8): 684–687. DOI: 10.1038/ng.628.

[24] Rivas MA, Beaun M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease[J]. Nat Genet, 2011, 43(11): 1066–1073. DOI: 10.1038/ng.952.

[25] Voight BF, Kang HM, Ding J, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits[J]. PLoS Genet, 2012, 8(8): e1002793. DOI: 10.1371/journal.pgen.1002793.

[26] Cortes A, Brown MA. Promise and pitfalls of the Immunochip[J]. Arthritis Res Ther, 2011, 13(1): 101. DOI: 10.1186/ar3204.

[27] Permuth JB, Pirie A, Ann Chen Y, et al. Exome genotyping arrays to identify rare and low frequency variants associated with epithelial ovarian cancer risk[J]. Hum Mol Genet, 2016, 25(16):

3600-3612. DOI:10.1093/hmg/ddw196.
[28] Grove ML, Yu B, Cochran BJ, et al. Best practices and joint calling of the HumanExome BeadChip: The CHARGE Consortium[J]. PLoS One, 2013, 8(7): e68095. DOI: 10.1371/journal.pone.0068095.
[29] Emond MJ, Louie T, Emerson J, et al. Exome sequencing of phenotypic extremes identifies CAV2 and TMC6 as interacting modifiers of chronic pseudomonas aeruginosa infection in cystic fibrosis[J]. PLoS Genet, 2015, 11(6): e1005424. DOI: 10.1371/journal.pgen.1005424.
[30] Lee S, Abecasis GR, Boehnke M, et al. Rare-variant association analysis: study designs and statistical tests[J]. Am J Hum Genet, 2014, 95(1): 5-23. DOI: 10.1016/j.ajhg.2014.06.009.
[31] 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)[J]. Mutat Res, 2007, 615(1-2): 28-56. DOI: 10.1016/j.mrfmmm.2006.09.003.
[32] Nicolae DL. Association tests for rare variants[J]. Annu Rev Genomics Hum Genet, 2016, 17: 117-130. DOI: 10.1146/annurev-genom-083115-022609.
[33] Price AL, Kryukov GV, de Bakker PIW, et al. Pooled association tests for rare variants in exon-resequencing studies[J]. Am J Hum Genet, 2010, 86(6): 832-838. DOI: 10.1016/j.ajhg.2010.04.005.
[34] Lange LA, Hu YN, Zhang H, et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol[J]. Am J Hum Genet, 2014, 94(2): 233-245. DOI: 10.1016/j.ajhg.2014.01.010.
[35] Victor RG, Haley RW, Willett DL, et al. The Dallas heart study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health[J]. Am J Cardiol, 2004, 93(12): 1473-1480. DOI: 10.1016/j.amjcard.2004.02.058.
[36] Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions[J]. PLoS Genet, 2010, 6(10): e1001156. DOI: 10.1371/journal.pgen.1001156.
[37] Firmann M, Mayor V, Vidal PM, et al. The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome[J]. BMC Cardiovasc Disord, 2008, 8: 6. DOI: 10.1186/1471-2261-8-6.
[38] Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies[J]. Am J Hum Genet, 2011, 89(3): 354-367. DOI: 10.1016/j.ajhg.2011.07.015.
[39] Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants[J]. Hum Hered, 2010, 70(1): 42-54. DOI: 10.1159/000288704.
[40] Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants[J]. PLoS One, 2010, 5(11): e13584. DOI: 10.1371/journal.pone.0013584.
[41] Ionita-Laza I, Buxbaum JD, Laird NM, et al. A new testing strategy to identify rare variants with either risk or protective effect on disease[J]. PLoS Genet, 2011, 7(2): e1001289. DOI: 10.1371/journal.pgen.1001289.
[42] Neale BM, Rivas MA, Voight BF, et al. Testing for an unusual distribution of rare variants[J]. PLoS Genet, 2011, 7(3): e1001322. DOI: 10.1371/journal.pgen.1001322.
[43] Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies[J]. Am J Hum Genet, 2010, 86(6): 929-942. DOI: 10.1016/j.ajhg.2010.05.002.
[44] Wu MC, Lee S, Cai TX, et al. Rare-variant association testing for sequencing data with the sequence kernel association test[J]. Am J Hum Genet, 2011, 89(1): 82-93. DOI: 10.1016/j.ajhg.2011.05.029.
[45] Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium[J]. Genet Epidemiol, 2009, 33(6): 497-507. DOI: 10.1002/gepi.20402.
[46] Faino A, Powell A, Williams A, et al. Identifying rare variants associated with hypertension using the C-alpha test[J]. BMC Proc, 2014, 8(S1): S56. DOI: 10.1186/1753-6561-8-S1-S56.
[47] Duchesne P, De Micheaux PL. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods[J]. Comput Stat Data an, 2010, 54(4): 858-862. DOI: 10.1016/j.csda.2009.11.025.
[48] Hirota K, Akagawa H, Onda H, et al. Association of Rare Nonsynonymous Variants in PKD1 and PKD2 with Familial Intracranial Aneurysms in a Japanese Population[J]. J Stroke Cerebrovasc Dis, 2016, 25(12): 2900-2906. DOI: 10.1016/j.jstrokecerebrovasdis.2016.08.002.
[49] Hasegawa T, Kojima K, Kawai Y, et al. AP-SKAT: highly-efficient genome-wide rare variant association test[J]. Bmc Genomics, 2016, 17(1): 745. DOI: 10.1186/s12864-016- 3094-3.
[50] Ll W, Li L, Nelson MR, et al. Deep resequencing unveils genetic architecture of ADIPOQ and identifies a novel low-frequency variant strongly associated with adiponectin variation[J]. Diabetes, 2012, 61(5): 1297-1301. DOI: 10.2337/db11-0985.
[51] Lin XY, Lee S, Wu MC, et al. Test for Rare Variants by Environment Interactions in Sequencing Association Studies[J]. Biometrics, 2016, 72(1): 156-164. DOI: 10.1111/biom.12368.
[52] 阮培峰. 家系数据中罕见基因变异与疾病关联分析的统计方法[J]. 复旦学报: 医学版, 2016, 43(2): 226-230. DOI: 10.3969/j.issn.1672-8467.2016.02.018.
Ruan PF. A statistical method for rare variants association studies in pedigree data[J]. Fudan Univ J Med Sci, 2016, 43(2): 226-230. DOI: 10.3969/j.issn.1672-8467.2016.02.018.
[53] Derkach A, Lawless JF, Sun L. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests[J]. Genet Epidemiol, 2013, 37(1): 110-121. DOI: 10.1002/gepi.21689.
[54] Fisher RA. Statistical methods for research workers[M]. London: Oliver and Boyd, 1925: 66-70.
[55] Lee S, Wu MC, Lin XH, et al. Optimal tests for rare variant effects in sequencing association studies[J]. Biostatistics, 2012, 13(4): 762-775. DOI: 10.1093/biostatistics/kxs014.
[56] Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies[J]. Am J Hum Genet, 2012, 91(2): 224-237. DOI: 10.1016/j.ajhg.2012.06.007.
[57] Lin PL, Tsai WY, Chung RH. A combined association test for rare variants using family and case-control data[J]. BMC Proc, 2016, 10(S7): 215-219. DOI: 10.1186/s12919-016-0033-x.
[58] Pan W, Kim J, Zhang YW, et al. A powerful and adaptive association test for rare variants[J]. Genetics, 2014, 197(4): 1081-1095. DOI: 10.1534/genetics.114.165035.
[59] Wei P, Cao Y, Zhang YW, et al. On Robust Association Testing for Quantitative Traits and Rare Variants[J]. G3 (Bethesda), 2016, 6(12): 3941-3950. DOI: 10.1534/g3.116.035485.
[60] Chen LS, Li H, Gamazon ER, et al. An exponential combination procedure for set-based association tests in sequencing studies[J]. Am J Hum Genet, 2012, 91(6): 977-986. DOI: 10.1016/j.ajhg.2012.09.017.
[61] Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60 706 humans[J]. Nature, 2016, 536(7616): 285-291. DOI: 10.1038/nature19057.

(收稿日期: 2016-12-28)
(本文编辑: 王岚)