基因关联分析中统计模型的回顾与教学思维*

张炳松　潘海燕　严海怡　陈逸懿　潘聪聪通讯作者

（广东医科大学公共卫生学院　广东东莞　523808）

摘　要：关联分析是筛选复杂疾病相关基因的重要手段，但要定位起因果关系的致病基因则需紧密结合计算科学与生物学知识。然而,在现实研究中，非统计学背景的研究者和学生往往难以理解关联分析的逻辑与结果，因此,常常造成统计学结论与生物学事实矛盾的窘境。文章将从关联分析的数据结构谈起，系统地回顾目前主流的基于广义线性模型的关联分析方法，侧重讨论各方法背后的统计学思维，旨在为相关领域研究者提供易于理解的统计学参考。

关键词：关联分析　单核苷酸多态性　罕见变异

中图分类号：G40-056　**文献标识码：**A

DOI：10.12218/j.issn.2095-4743.2023.01.164

二十一世纪以来，随着生命科学和计算机技术等的飞速发展，复杂疾病的诊断和治疗研究进入了系统和生物组学的新时代。复杂疾病也称为多基因病，其发生与发展涉及多种基因和环境刺激。现代医学认为，诱发复杂疾病的内因是遗传物质的改变，如基因突变、单核苷酸多态性等，这些改变可以直接导致机体功能先天性损伤，或致使机体对环境产生后天性易感[1]。环境与遗传对复杂疾病的成因贡献比例并无定论，但一般认为除非是机体暴露于极端的环境中（如核辐射等），否则遗传物质是复杂疾病形成与发展的核心因素，因此，比较遗传物质的差异即成为当今研究复杂疾病重要的手段之一。

人类的DNA序列中大约有0.2%的差异，即任取人群中的两个独立个体，其体内99.8%的序列信息是一致的。这微小的序列差异有不同的类型和作用形式，其中最常见的是单核苷酸多态性（SNP），或称为变异。根据在人群中少数等位基因频率（MAF）的高低，变异可以分为常见变异和罕见变异（MAF低于1%），探索变异与疾病之间的关系的统计学方法即为关联分析。对于复杂疾病的研究，关联分析仅能从统计学角度筛选出潜在的与疾病关联的风险变异，要实现关联关系到因果关系的转换，则需要结合计算科学家与生物学家等的努力。然而，在现实的科研合作和教育教学中，非统计学研究者往往难以理解关联分析的构建思路，也无法区分各关联分析的差别与应用场景，甚至对关联分析研究中的数据结构一无所知。鉴于此，本研究将从关联分析研究的数据结构谈起，创新性地从统计思维的角度对目前主流的基于广义线性模型的关联分析方法做系统的回顾，旨在给非统计专业的研究者梳理关联分析研究的框架，提供易于理解的参考，帮助其从根本上理解各方法之间的联系与差异，为日后的合作打下良好的基础。

一、材料和方法

基因变异在复杂疾病的发病过程中起到重要的作用，但其作用机理尚不明确。其中，主流的变异-复杂疾病关联假说有“常见疾病-常见变异假说”和“多重罕见变异假说”[2]。前者认为复杂疾病是由多个常见变异共同作用的结果，每个变异的作用是微效的，而累积效应则引起疾病的发生。后者则认为疾病的发生取决于某些特定的效应较大的变异。根据这两套理论，统计学的关联分析可以分为单位点分析和多位点分析，分别对应常见变异筛选的分析和罕见变异筛选。

1．材料

单位点分析与多位点分析均基于高通量测序数据中SNP的合集，图1和图2展示了数据的生成过程。在人群序列中，若某一位点上的基因分型不唯一，则称在此位点上出现单核苷酸多态性。如图1所示，个体1与个体2的第四号位点上基因型不同，则该位点为SNP位点，而其余位点为非SNP位点。在获得个体的全序列数据后，关联分析研究只收集SNP位点的信息，并根据少数等位基因数目将其碱基信息转化为数值信息。如图2所示，假设第2号位点为SNP位点，该位点的多数等位基因为C，则该个体在此位点上的少数基因数目为0。关联分析研究常用的数据集即是人口学信息与SNP信息的集合，人口学信息包括性别、年龄、种族等，而SNP信息则为各位点上少数等位基因的数目。

*基金项目：教育部人文社会科研研究一般项目（NO:21YJC910007）；广东省普通高校重点平台重点领域专项（NO:2020ZDZX3007）；东莞市科技特派员项目（NO:20201800500082）；广东省普通高校青年创新人才类项目（NO:2022KQNCX021）。

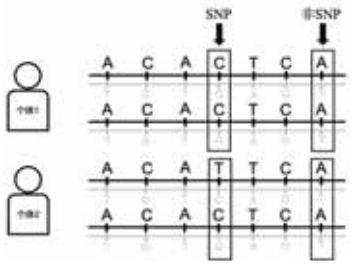

图 1　单核苷酸多态性（SNP）示意图

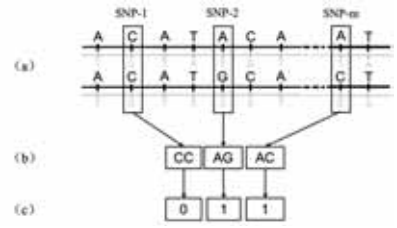

图2　基因关联分析研究中SNP数据产生流程

2．单位点分析

单位点分析即分析单一位点与性状（疾病发生）之间的关联关系，常用于分析常见变异与疾病之间的关系。设用于关联分析的样本里包含 n 个独立的个体，第 i 号个体的患病情况为 y_i，其对应人口学信息（如年龄、性别等）可用 $x_{i1},x_{i2},\cdots,x_{ip}$ 等来表示。另设序列片段包含 m 个突变位点，其中 g_{ij} 表示第 i 号个体的第 j 号位点上少数等位基因的个数（$g_{ij}=0,1,2$），则该个体的患病情况与突变位点关联关系可用以下模型来描述：

$$g(y_i)=\alpha_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}+\gamma g_{ij},\#(1)$$

其中，$g(\cdot)$ 为连接函数（对于如是否患病的二分类变量，连接函数通常为logit函数），α_0 为基线患病风险，$\beta_1,\cdots,\beta_p$ 为对应人口学变量对患病的贡献，γ 为第 j 号位点的变异对患病的贡献。检验位点与患病风险的关联即检验 H_0：$\gamma=0$，这与检验其他变量的显著性并无差异。模型（1）也称为线性加性模型，其特征是各变量之间等权加和，且对于每一个变量而言，其增量对患病的影响是恒定。该模型的线性关系意味着性别、年龄等宏观的因素与基因等微观因素对患病的影响是可比的（数量级相当的），任何一方都不在关系中占绝对优势；而加性关系则表示少数等位基因的个数对患病风险的影响是恒定的。

单位点分析是早期全基因组关联分析的核心模型之一，对单基因缺陷遗传病风险基因的探测尤为重要。但复杂疾病的致病机制往往涉及多个变异，单独分析变异的做法会忽视其联合作用，因此多位点模型应运而生。

3．多位点分析

除了上述提及的复杂疾病的致病机理催生多位点分析外，最小等位基因频率过低亦是重要的考量。若无所需量级的样本，则检验效能将低于预设水平。然而，这样本量不但远超绝大多数高通量关联研究，而且也对计算设备带来极大的挑战，因此，单位点的分析难以筛选出罕见变异。统计学上多位点的分析方法众多，但构造理念并不多，以下将从核心统计思维的角度介绍常用的三种。

（1）负荷检验（Burden test）

负荷检验的核心思想源于统计学上常规地对罕见事件的处理手段：合并罕见事件[3]。承接模型（1）的数学符号与设定，若 $g_{i1},g_{i2},\cdots,g_{ig}$ 为个体 i 的 q 个位点上的少数等位基因频数，则其与疾病的关联可用以下模型描述：

$$g(y_i)=\alpha_0+\beta_1 x_{i1}+\cdots\beta_p x_{ip}+\gamma\sum_{j=1}^{q}g_{ij\cdot}\#(2)\ a$$

该模型把序列片段内所有罕见变异等权线性合并组成一个新的“变异”，并将其与宏观因素共同关联疾病的发生。检验序列片段与患病风险的关联即检验 H_0：$\gamma=0$，这与检验其他宏观变量的显著性并无差异。负荷检验的做法易于理解且运算便捷，因此，成为应用最广的罕见变异分析方法之一。然而，从统计学上看，该模型基于两个不易察觉的且违反常理的假设。将模型（2）改写成如下形式：

$$g(y_i)=\alpha_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}+\gamma g_{i1}+\cdots+\gamma g_{iq\cdot}\#(3)$$

对于某一序列片段，负荷检验模型假设片段上的所有位点皆为致病位点，且各位点对患病风险的贡献皆相等且方向一致。在复杂疾病的研究当中，学界的主流共识是片段上的位点对疾病的发生贡献不一。有研究验证，负荷检验在混合效应的片段上检验效能较其他方法低，但在单独效应的片段上检验效能最高[9]。在实际操作中，由于无法得知所测片段上变异对疾病的影响方向，因此，处于运算速度考虑，研究者会优先使用负荷检验筛选片段，所得显著序列则可能为单独效应的序列。

2．序列核函数关联性检验（Sequence-kernel association test，SKAT）

为了克服负荷检验中违反生物学常规的假设，SKAT跳出加性模型的统计学框架，用混合效应模型的思想重新构建变异与疾病的相关关系[4]。承接模型（2）的数学符号与设定，SKAT用以下模型描述序列片段与患病风险的关系：

$$g(y_i)=\alpha_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}+\gamma_1 g_{i1}+\cdots+\gamma_q g_{iq\cdot}\#(4)$$

其中$\gamma_j \sim N(\gamma_0, w_j\tau), j=1,\cdots,q$。检验序列片段与患病风险的关联即检验$H_0$：$\gamma_0=0$，或者$H_0$：$\tau=0$。这背后的统计学思想是，假设随机变量$V$服从均值为，方差为0的正态分布，则$V=c$，即$V$是没有变化的固定效应或常数。与模型（3）相比，模型（4）允许序列片段上各位点对疾病的贡献效应与影响方向不定，这使得SKAT更适用于复杂疾病的风险基因筛选。除此以外，模型（4）将序列上位点间的关联用权重参数拟合，克服了负荷检验中独立位点假设的缺陷。从统计学的角度看，该模型的构建满足节俭原则，即构建“最简单的可信模型”。当对w_j作适当分布假设时，仅需通过少量分布参数估计即可把各位点在序列上的权重悉数估计，这在位点数量巨大的序列上优势尤其明显。有研究验证，当序列上存在正向与反向作用的变异时，使用SKAT较负荷检验有更高更稳定的统计性能。在实际操作中，研究者常常会使用SKAT重新筛查所有基因，而事实上负荷检验筛选的基因大概率能通过SKAT，尤其是p值非常小的片段。因此，全基因组关联分析的第二阶段仅需对负荷检验边缘显著或不显著的片段进行筛选即可。

SKAT克服了负荷检验中不合常规的假设问题，同时有不错的统计效能与运算速度，因此，成为目前罕见变异关联分析的主流方法之一。然而，负荷检验与SKAT仅考虑片段上位点的信息，并未考虑位点所在位置的信息，即连锁不平衡的影响，因此，有FRT的提出来解决此问题。

3．函数型回归否检验（functional-based regression test，FRT）

为了兼顾位点与位置信息对疾病风险的影响，现考虑以下的模型构建框架[5]。假设关联分析研究中收集到n个个体m个位点的少数等位基因数量的信息，$x_{i1}, x_{i2}, \cdots, x_{ip}$对应$p$个人口学等因素，$g_i(u)$为第$u$号位点上的少数等位基因数量，则序列片段与患病风险的关联可用以下模型描述：

$$g(y_i)=\alpha_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}+\sum_{u=1}^{m}\gamma(u)g_i(u).\#(5)$$

其中，$\gamma(u)$为第u号位点对患病风险的效应量，是一个关于序列上位点位置的函数。检验序列片段与患病风险的关联即检验H_0：$\gamma(u)=0$。与模型（4）相比，模型（5）在允许位点的效应量与效应方向各不相同的基础上，通过假设效应量函数把位点的相对位置信息储存于连续变化的关系当中，因此较SKAT更贴近生物学事实。从统计学角度看，模型（1）（2）（4）均把位点看成独立随机变量与患病风险构建统计模型，而模型（5）则把一系列随机变量看成随机过程构建模型，因此拟合了各位点间相互关系的信息。除此以外，由于变异的效应量存储在连续函数$\gamma(u)$中，序列上少量的缺失信息不会对函数造成本质的改变，因此，FRT在部分缺失的序列数据上有稳健的表现，这是负荷检验与SKAT不具备的优势。有研究验证，在混合效应的序列片段上，FRT的效能要优于其他两种检验，而在单独效应片段上的检验效能与负荷检验相当。

二、讨论

基因的关联分析在过去的十余年间取得了长足发展，从单位点分析到多位点联合分析，从固定效应模型到混合效应模型再到函数型回归模型。本文从统计模型构建的思维入手，回顾了几种基于广义线性模型的关联分析方法，旨在帮助非统计专业的研究人员和学生快速入门，从本质上理解各关联分析模型间的区别。负荷检验通过合并罕见变异的做法提高检验效能，其缺点是该模型的假设与生物学常规不符。基于负荷检验，后续有通过权重来抵消其假设的方法，统称为负荷系检验[6]。SKAT从混合效应模型的思维中对关联分析模型进行重构，它不再假设变异的效应量为固定效应，而是服从正态分布的随机效应。通过检验其随机性而非效应量本身去衡量与患病的关联关系。负荷检验与SKAT均有其优势的使用情景，后续还有基于权重的融合负荷检验与SKAT的模型，统称为SKAT系检验[7]。负荷系与SKAT系理论完整，软件完善，是目前主流的罕见变异筛选方法，但两者均未考虑连锁不平衡的影响。FRT的提出首次解决这个问题，通过引进效应量函数把位点信息与位置统一到一个模型里。它的优点是更完整地描述了基因片段与患病风险的关联关系，但缺点是计算速度慢，且理论层面尚缺快速检验的统计量构建方法。从统计思维的角度看，三种模型均基于广义线性模型的框架。这个框架理论成熟，扩展性强，可解释性高，软件完善，是目前主流的分析方法。除此以外，还有合并p值、分布拟合等关联分析框架[8]，但这些方法的回顾超过了本文范围，因此不在此讨论。

对于非统计专业背景的研究者，区分本文提及的三种关联分析方法的关键是熟悉各方法的假设条件。对于某一段基因序列，负荷检验假设各位点均为致病位点，且其对疾病的贡献均相等；序列核函数关联性检验假设各位点的效应不同，但每个位点对疾病的效应因人而异，不同位点的效应在人群中的离散程度成比例；函数型回归检验则假设各位点的效应构成连续变化的函数，根据其位置决定对疾病的贡献。另外，基因关联分析的结果需要校正检验水平以避免假阳

性；关联基因亦需通过结合生物学知识以决定其因果效应。

医学院校的本科教育中，非医学统计学专业学生，如生物技术等，大多通过《生物信息学》课程初次接触基因关联分析。由于专业偏向与课程设计的缘故，该专业学生尚不具备数据和统计学相关思维，在教学设计中建议从这两方面的基本知识入手。数据库的基本知识包括二维表及关联关系，及基本表格包含行与列（分别对应统计学的记录与变量），不同表格之间通过主键来联系。统计学的基本知识包括总体和样本的概念，假设检验，回归的思想与实现，统计结果的解释等。基因关联分析的入门软件是Plink，一款基于Linux操作系统的分析软件，因此学生在掌握理论知识的基础上仍需实验课的联系以全面了解关联分析的操作流程。除了数据格式和数据清理的操作讲授外，实验课应该涉及全数据关联分析、亚组关联分析等，统计模型则包括等位基因关联分析、加性模型、显性模型、隐性模型、一般模型等。课堂练习则以代码逻辑和结果解释为主，课后作业则可设计需要自主思考的小分析项目。在教学实践中发现，学生经过理论与实验的讲授，能完成数据录入、数据清理、运行计算、结果解释的全流程，也能基本判断其他关联分析流程的完整性。

关联分析的应用对复杂疾病的研究至关重要。它通过统计学的方法定位与疾病相关联的风险变异，已成功应用于多种重大疾病。然而，关联分析也遇到很大的质疑，从最开始的太少的关联变异到现在太多的关联变异，但遗传率的解释方面并没有本质提升，因此，研究者慢慢把目光转移到丢失遗传机制的研究上[12]。目前的关联分析方法从变异筛选的角度切入，但对遗传机制的关注相对薄弱，仅有FRT考虑到变异位置之间的联系，但也仅考虑相对位置而非绝对物理距离。综合来看，基因的关联分析已经为复杂疾病的研究积累大量的研究成果，但未来依然有很大的发展空间。复杂疾病的研究仍需生物统计学家、遗传学家、分子生物学家、药学家等通力合作，希望本文能为非统计学专业的学者们提供易于理解的参考，为未来的合作奠定坚实的基础。

参考文献

[1]Florez JC, Hirschhorn J, Altshuler D. The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits[J]. Annu Rev Genomics Hum Genet, 2003,4:257-91.

[2]李霞，雷健波，李亦学. 生物信息学（第2版）[M]. 北京：人民卫生出版社, 2015.6.

[3]Li, B., and Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data[J]. Am. J. Hum. Genet,2008(83): 311–321.

[4]Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test[J]. Am J Hum Genet, 2011(89):82-93.

[5]Fan, R., Zhang, Y., Albert, P., Liu, A., Wang, Y., and Xiong, M. Longitudinal association analysis of quantitative traits[J]. Genet Epidemiol，2012，36:856-869.

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

[7]Price, A.L., Kryukov, G.V., de Bakker, P.I.W., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. Pooled association tests for rare variants in exon-resequencing studies[J]. Am. J.Hum. Genet，2010(86)：832–838.

[8] Zhang, B., Chiu, C., Yuan, F., Sang, T., Cook, R., Wilson, A., Bailey-Wilson, J., Chew, E., Xiong, M., Fan, R. Gene-based analysis of bi-variate survival traits via functional regressions with applications to eye disease[J]. Genet Epidemiol. 2021. 45(5):455-470.

[9] Zhang, B., Wang, S., Mei, X., Han, Y., Wang, R., Fang, H., Chiu, C., Ding, J., Wang, Z., Wilson, A., Bailey-Wilson, J., Xiong, M., Fan, R. Stochastic functional linear models for gene-based association analysis of quantitative traits in longitudinal studies[J]. Statistics and Its Interface. 2022. 15:181-196.

[10]Ionita-Laza, I., Buxbaum, J.D., Laird, N.M., and Lange, C. A new testing strategy to identify rare variants with either risk or protective effect on disease[J]. PLoS Genet，2011(7).

[11] Lee, S., Wu, M.C., and Lin, X. Optimal tests for rare variant effects in sequencing association studies[J]. Biostatistics 2012(13), 762–775.

[12] Yang, J., Benyamin, B., McEvoy, B. et al. Common SNPs explain a large proportion of the heritability for human height[J]. Nat Genet. 2010(42), 565–569.

通讯作者

潘聪聪，女，公共卫生学院（养老产业院），讲师。