Several Problems Appearing in Genome-Wide Association Study

PENG Qianqian
MOE Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433 China

ABSTRACT: The coming-forth of genome-wide association study had brought in new frontiers of complex disease association study, which was designed to discover all the disease loci. But along with the processing of genome-wide association study, the result of it was not as satisfying as researchers primarily expected. But the imperfect outcome of genome-wide association study was not only due to the methods applied to it, but there were also many other reasons. In this review, we focused on some problems encountered in genome-wide association study, including the argument among statistical interaction and biological interaction, the characteristics of genome-wide genotyping data and related methods developing, hypothesis of disease model and so on. The renewed learning and thinking on the genome-wide association study would give efficient supervision in the following study.

Keywords: genome-wide association study; interaction; characteristics of data; model hypothesis

对全基因组关联分析中一些问题的思考

彭倩倩
复旦大学生命科学学院现代人类学教育部重点实验室 上海 200433

摘要：全基因组关联分析的出现，曾经令研究者们看到了复杂疾病致病机制研究的新契机——从全基因组分型数据中寻找所有可能的致病位点。但是随着全基因组关联分析的逐步推进，其结果并没有研究者预期的那么理想。全基因组关联分析的结果不理想，不仅仅是分析方法的问题。本文对全基因组关联分析中遇到的一些问题，如交互作用的争议，全基因组分型数据的特征以及相关方法构建，疾病模型假设等方面进行了讨论。对于全基因组关联分析的重新认识和思考对今后的工作将提供有力的指导。

关键字：全基因组关联分析；交互作用；数据特征；模型假设

全基因组关联分析的出现，曾经令研究者们看到了复杂疾病致病机制研究的新契机——从全基因组分型数据中寻找所有可能的致病位点，而不必逐个寻找候选基因然后用定位方法确定[1-5]。但是随着全基因组关联分析的逐步推进，其结果并没有研究者预期的那么理想，引起了研究者们一系列的反思[6-11]。首先在数据分析方法上作了一系列改进和探索，大量应用于全基因组关联分析的方法涌现，例如基于数据挖掘的方法[12-19]、基于连锁不平衡或熵的方法[20-24]，基于整体基因的方法[25-26]，基于交互作用的方法[27-29]等。但是这些方法并没有在全基因组关联分析中得到很好的应用，究其原因一方面是新方法还没有普及使用，另一方面是应用效果改善不大制约了其推广。但是全基因组关联分析结果不佳，也并非单纯是分析方法的问题，实验设计、数据特征以及模型假设等也是不可忽视的问题。本文对研究中遇到的一些有关全基因组关联分析的问题做一点总结探讨。

1 揭秘交互作用的玄虚

1.1 交互作用在生物学领域的意义

一般交互作用是这样定义的，两个因素同时作用时的效应与其边际效应乘积的偏离称之为两个因子的交互作用[30]。通俗的解释就是两个因素在效应上的修饰性[31]。可见，交互作用的定义是建立在表型的基础上的。但是交互作用又是怎么形成的呢？从生物学的角度来讲，基因水平的变异(单碱基突变，多态以及大片段的重复、插入、缺失等)导致分子水平的改变，从而最终出现表型上的差异。基因与基因之间的交互作用实际上是基因与基因的表达产物之间的关系(例如同一通路上的共同调控、上下游或替代效应)出现了差异，从而导致表型变异。因此基因与基因之间的交互作用是直接与基因的功能水平相关的。

收稿日期：2010 年 6 月 7 日 **修回日期**：2010 年 6 月 11 日 **联系人**：彭倩倩 qqpeng2009@gmail.com

1.2 统计学中对交互作用的定义

下面以 logistic 回归模型对交互作用的研究为例探讨交互作用的统计学阐释[30]。研究两个位点(SNP)的交互作用，设 y 表示疾病状态，两个 SNP 分别用 x1 和 x2 表示，

$$\log it(y) = \alpha + \beta 1x1 + \beta 2x2 + \gamma x1x2$$

x1 和 x2 的基底效应表示为 e^{α}， x1 的效应表示为 $e^{\alpha+\beta 1}$， x2 的效应表示为 $e^{\alpha+\beta 2}$， x1 和 x2 的效应表示为 $e^{\alpha+\beta 1+\beta 2+\gamma}$。

这里面首先应该扣除基底效应 e^{α}，那么交互作用的效应

$$Effect_{x1x2} = \frac{e^{\alpha+\beta 1+\beta 2+\gamma}}{e^{\alpha}} - \frac{e^{\alpha+\beta 1}}{e^{\alpha}} \cdot \frac{e^{\alpha+\beta 2}}{e^{\alpha}},$$

也即 $Effect_{x1x2} = e^{\beta 1} \cdot e^{\beta 2}(e^{\gamma} - 1)$。

在交互作用为 0 的假设下，

$Effect_{x1x2} = e^{\beta 1} \cdot e^{\beta 2}(e^{\gamma} - 1) = 0 \Rightarrow \gamma = 0$。

可见，统计学中对交互作用的度量是通过不独立性程度度量，

$$P = P(AB) - P(A)P(B),$$

即两个基因的协同效应是否等于其边际效应的乘积。

理论上讲，对照组的交互作用应该为 0，但是由于复杂的原因使得对照组中交互作用效应不呈现为 0。因此在构造统计量时要扣除这部分效应。对于其他的统计量，如基于 LD 或基于熵的统计量以及基于典型相关的统计量，交互作用的表现形式也是

$P = P(AB) - P(A)P(B)$。

1.3 统计学交互作用与生物学交互作用的争议

统计学意义上的交互作用的原理已经整理清楚，但是历来已久的统计学意义的交互作用与生物学意义的交互作用的争论又作何解释呢[25, 28, 31]？考虑一个简单的情况，A 因素与 B 因素存在交互作用，B 因素与 C 因素存在交互作用，那么 A 因素与 C 因素是否存在交互作用？根据统计学意义上的交互作用定义，很有可能得到的结论是 A 因素与 C 因素存在交互作用，但是生物学则要怀疑表面上的 A 因素与 C 因素之间的交互作用可能是由 A 因素与 B 因素的交互作用以及 B 因素与 C 因素的交互作用造成的，实际上 A 因素与 C 因素之间不存在交互作用。单纯统计学上得到的两两因素之间的交互作用检验结果并不能否认这个可能性。因此，统计学意义的交互作用与生物学意义的交互作用的分歧就在这里，统计学仅仅找出了“可能”是交互作用的结果，而不能回答是不是因素之间真实存在的交互作用。将统计学意义的交互作用阐述的再好，我们也回答不了生物学意义的交互作用，因为我们不能穷尽所有的因素并一一排除他们的影响。从统计学意义的交互作用本身来论证生物学意义的交互作用是徒劳无功的。

要解决统计学意义的交互作用与生物学意义的交互作用的分歧，可以参考基因表达芯片数据的分析方法[32-33]。对于表达谱芯片数据，一般是这样处理的：如果有两个因素显示是非独立的，在考虑所有剩余因素的影响后仍是非独立的，那么他们是不独立的，否则两个因素是独立的，或者简化一下，如果两个因素在依次考虑了另外一个剩余因素的影响后其结果都是不独立的，那么这两个因素可以认为是非独立的。对于全基因组关联分析数据的交互作用的研究也可以借鉴这种思路，但是对于全基因组关联分析数据，与基因表达芯片数据存在共同的问题，就是根据此方法构建的网络的可读性非常差，特别对于基于全基因组分型的 SNP 数据构建的网络，其应用性较弱[34-35]。

2 全基因组分型数据的特征

全基因组分型数据是在全基因组范围内，根据“单倍域”理论以及连锁不平衡的原理随机或者有偏向性的铺点，根据分型的 SNP 来定位真正的致病位点的位置。全基因组分型数据有几个假设：致病基因与分型的 SNP 位点位于同一个“单倍域”中；致病基因与分型的 SNP 位点的连锁不平衡程度很高；致病基因在总群体中的频率不是很低。这三条是全基因组关联分析的工作原理，特别是当致病基因与 SNP 的连锁很紧密的时候，此时才最有可能检出致病基因。但是我们并不知道真正的致病基因在哪里，因此我们只能期望全基因组关联分析所检测出来的显著位点附近隐藏着真正的致病基因。但是迄今为止，全基因组关联分析的结果被重复出来的

比例很小[7, 9]。

3 基于整体基因的分析方法

针对这个问题，近年来很多研究者作了许多探讨。除了基于单个 SNP 的关联分析方法的改进，还有基于整体基因[25-26]，基于交互作用[12, 21, 24, 27]以及基于通路研究[34-35]的发展。基于整体基因的研究是在全基因组分型数据的基础上，将位于一个基因区域内的 SNP 先按照一定的方法综合信息，而后以基因为单位进行关联分析。该方法一方面充分利用了 SNP 的信息，降低了关联分析的复杂度，另一方面将基因作为一个功能单位，而不是以 SNP 为单位，其生物学意义更为明显。整体基因是介于单个 SNP 与 haplotype 之间的一个研究变量：与 SNP 相比，其可以更好的代表基因的信息；与 haplotype 相比，其省去了单体型推断带来的一系列问题，而且可以有效的降低维度[25-26]。另一方面，基于整体基因的研究更实用，而且研究结果对个体基因型的解释相对更容易，毕竟实验得到的第一手数据是基因型数据。

对于整体基因的信息如何综合是重要问题。众所周知，基因内的 SNP 越多，那么其隐含的基因的信息肯定越大。目前采用的整体基因信息综合的方法主要有主成分分析和典型相关分析。

基于主成分或基于典型相关的整体基因分析方法只能在已有的 SNP 分型数据的基础上将病例组与对照组的差异寻找出来，而这些差异是否与疾病相关，其解释能力有几许尚未可知？这种纯线性的数据变换有时候可以将真正的致病效应湮没。例如，当基因区域内只有一个 SNP 与致病基因关联且效应显著时，此时利用基于主成分分析的整体基因方法可能并不能检测出整体基因与疾病的关联性。因此，在将其应用于大规模关联分析数据时对结果解释要慎重。

基因分型数据与表达数据不同，表达数据直接代表了整个基因的信息，其所看的是表达量与所研究性状的关联性。而分型数据与基因的关系不似基因表达数据那样直观，SNP 与致病基因之间存在连锁不平衡的关系。近来有研究表明，即使 SNP 的检验结果是显著的，其所反映的致病基因的位置可能离该 SNP 非常远[9]。这就使得 SNP 与致病基因之间的关系更加难以解释。标签 SNP 的本意是通过其与致病基因的连锁不平衡，利用标签 SNP 来寻找致病基因或致病位点。但是基于这些标签 SNP 来整合基因的变异信息是否合理或如何整合这个信息还有待商榷。

主成分分析或典型相关分析等基于降维的方法在群体研究中有很多优点，因为群体之间的离散信息可以集中在几个维度上考察。但是对于疾病研究来说，仅仅有这些离散信息不够，还要进一步的去探索这些离散信息的源头以及离散信息与疾病表型之间的关系。一般的理解，这些离散信息代表了疾病表型与对照的差异，然而这些信息的源头又难以对应到特定的 SNP 上，这与标签 SNP 的定义的初衷相悖。主成分分析等降维方法将原始数据映射到特征空间中，但是该特征空间并没有考虑是否能有效的分割病例对照，该问题是基于整体基因关联分析的主要弊病。虽然有 control-based 方法可以在一定程度上改进，但是效果并不十分明显[26]。整体基因是一个很好的概念，但是在对整体基因信息的整合上还存在着一些问题，需要进一步改进。

4 疾病模型假设的问题

全基因组分型数据在分析上困难重重，除了方法的局限性，也可能是本身的模型假设和实验设计就存在问题。对于复杂疾病关联研究，一般的假设是“常见疾病-常见变异(common disease-common variation)”。在研究相对高发疾病和主基因效应明显的疾病时，该假设是有效的。复杂疾病的发病率虽然一般较高，但是其致病机制却错综复杂，通常不能由几个主效基因解释。另外一种猜想是疾病群体本身存在异质性，假设“少数病人-几个主效基因”的致病模型成立，在整个群体中的表现就是“常见疾病-多个微效基因(common disease-multiple rare variations)” [7, 9]。这种猜想直接对全基因组分型数据的实验设计进行了否定，因为全基因组分型数据的一个要求就是最小等位基因频率(MAF)>0.05。承认“常见疾病-多个微效变异”模型，那么全基因组关联分析的效力就大打折扣。据此，有人通过模拟研究了在该模型

的假设下全基因组关联分析的效力。结果表明全基因组分型数据可以检出 30%左右的由相对低频等位基因造成的致病效应，同时表明所检出的 SNP 所代表的致病基因可能离该 SNP 较远[7, 9]。反言之，如果这个模型正确，那么利用现今通用的方法通过全基因组关联分析直接寻找真正的致病基因几乎是不可行的。

复杂疾病致病通路的复杂性是目前研究的主要障碍，发展合理、有效的疾病模型，以及有效的切入研究方法显得尤为重要[7-10]。不同的致病模型假设也需要发展不同的分析方法。

总而言之，现在全基因组关联分析存在着诸多问题亟待解决。对于全基因组关联分析的各个方面的反思，从疾病模型假设到实验设计，从分析方法到结果验证，可以使我们客观的认识全基因组关联分析的优点与不足。这些新的认识将可以指导新的实验设计制定以及构建更加高效的分析方法，提高全基因组关联分析的检验效力。

参考文献

1.McCarthy MI, Zeggini E (2009) Genome-wide association studies in type 2 diabetes. Curr Diab Rep 9: 164-171.

2.Hirschhorn JN, Lettre G (2009) Progress in genome-wide association studies of human height. Horm Res 71 Suppl 2: 5-13.

3.Grant SF, Hakonarson H (2009) Genome-wide association studies in type 1 diabetes. Curr Diab Rep 9: 157-163.

4.Franke B, Neale BM, Faraone SV (2009) Genome-wide association studies in ADHD. Hum Genet 126: 13-50.

5.Ozaki K, Tanaka T (2005) Genome-wide association study to identify SNPs conferring risk of myocardial infarction and their functional analyses. Cell Mol Life Sci 62: 1804-1813.

6.Weale ME (2010) Quality control for genome-wide association studies. Methods Mol Biol 628: 341-372.

7.Robinson R (2010) Common disease, multiple rare (and distant) variants. PLoS Biol 8: e1000293.

8.Need AC, Goldstein DB (2010) Whole genome association studies in complex diseases: where do we stand? Dialogues Clin Neurosci 12: 37-46.

9.Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8: e1000294.

10.Cho JH (2010) Genome-wide association studies: present status and future directions. Gastroenterology 138: 1668-1672 e1661.

11.Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-9367.

12.Garcia-Magarinos M, Lopez-de-Ullibarri I, Cao R, Salas A (2009) Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction. Ann Hum Genet 73: 360-369.

13.Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol 241: 252-261.

14.Brassat D, Motsinger AA, Caillier SJ, Erlich HA, Walker K, Steiner LL, Cree BA, Barcellos LF, Pericak-Vance MA, Schmidt S, Gregory S, Hauser SL, Haines JL, Oksenberg JR, Ritchie MD (2006) Multifactor dimensionality reduction reveals gene-gene interactions associated with multiple sclerosis susceptibility in African Americans. Genes Immun 7: 310-315.

15.Cook NR, Zee RY, Ridker PM (2004) Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. Stat Med 23: 1439-1453.

16.Bastone L, Reilly M, Rader DJ, Foulkes AS (2004) MDR and PRP: a comparison of methods for high-order genotype-phenotype associations. Hum Hered 58: 82-92.

17.Hahn LW, Ritchie MD, Moore JH (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19: 376-382.

18.Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69: 138-147.

19.Nelson MR, Kardia SL, Ferrell RE, Sing CF (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 11: 458-470.

20.Wu X, Jin L, Xiong M (2008) Composite measure of linkage disequilibrium for testing interaction between unlinked loci. Eur J Hum Genet 16: 644-651.

21.Kang G, Yue W, Zhang J, Cui Y, Zuo Y, Zhang D (2008) An entropy-based approach for testing genetic epistasis underlying complex diseases. J Theor Biol 250: 362-374.

22.Dong C, Chu X, Wang Y, Jin L, Shi T, Huang W, Li Y (2008) Exploration of gene-gene interaction effects using entropy-based methods. Eur J Hum Genet 16: 229-235.

23.Cui Y, Kang G, Sun K, Qian M, Romero R, Fu W (2008) Gene-centric genomewide association study via entropy. Genetics 179: 637-650.

24.Zhao J, Jin L, Xiong M (2006) Test for interaction between two unlinked loci. Am J Hum Genet 79: 831-845.

25.Peng Q, Zhao J, Xue F (2010) A gene-based method for detecting gene-gene co-association in a case-control association study. Eur J Hum Genet 18: 582-587.

26.Peng Q, Zhao J, Xue F (2010) PCA-based bootstrap confidence interval tests for gene-disease association involving multiple SNPs. BMC Genet 11: 6.

27.Jung J, Zhao Y (2010) Allelic based gene-gene interaction in case-control studies. Hum Hered 69: 14-27.

28.Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10: 392-404.

29.Musani SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari HK, Allison DB (2007) Detection of gene x gene interactions in genome-wide association studies of human population data. Hum Hered 63: 67-84.

30.贾崇奇(2005) 交互作用分析方法. 见: 赵仲堂(ed) 流行病学研究方法与应用. 北京: 科学出版社, 523-544.

31.Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11: 2463-2468.

32.Gronlund A, Bhalerao RP, Karlsson J (2009) Modular gene expression in Poplar: a multilayer network approach. New Phytol 181: 315-322.

33.Horvath S, Dong J (2008) Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 4: e1000117.

34.Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, Amos CI, Xiong M (2010) Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet 18: 111-117.

35.Luo L, Peng G, Zhu Y, Dong H, Amos CI, Xiong M (2010) Genome-wide gene and pathway analysis. Eur J Hum Genet, in press.