基因大数据的集成分析

胡湘红[1]，彭衡[2]，杨灿[3]，张纵辉[1]，万翔[1]，罗智泉[1]
1. 深圳市大数据研究院，广东 深圳 518172；2. 香港浸会大学数学系，香港 999077；
3. 香港科技大学数学系，香港 999077

摘要
随着生物科技（如基因芯片和测序技术）的飞速发展，全世界已经积累了海量的数据。有效地整合和集成多层面和多维度的基因大数据，对于全方位解析从遗传变异到疾病发生的整个因果链条具有关键作用，可为个性化、精准医疗服务奠定科学的基础。从3个方面对基因大数据的集成分析进行综述：检测风险位点及其功能分析、基因多效性的分析、基于孟德尔随机化的因果推断。进一步结合具体的应用案例进行了阐述，最后对基因大数据的集成分析研究进行了总结以及展望。

关键词
GWAS；集成分析；多基因效应；基因多效性；孟德尔随机化

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2019033

Integrative analysis for big data in genomics

HU Xianghong[1], PENG Heng[2], YANG Can[3], CHANG Tsunghui[1], WAN Xiang[1], LUO Zhiquan[1]
1. Shenzhen Research Institute of Big Data, Shenzhen 518172, China
2. Department of Mathematics, Hong Kong Baptist University, Hong Kong 999077, China
3. Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong 999077, China

Abstract
With the rapid development of bio-technology (e.g., genotyping chip and sequencing), world-wide researchers have accumulated massive data sets at different levels. Integrative analysis of multi-layered genomic data can greatly contribute to the completion of causal chain from genetic variants to phenotype variations, laying a scientific foundation for personalized and precise medicine. The integrative analysis from the following three aspects mainly reviewed: identification of causal variants and their functional annotation, pleiotropy in human complex traits, Mendelian randomization forcausal inference between phenotypes, and several case studies were provided. Finally, the importance of integrative analysis in genomic data for precision medicine was highlighted.

Key words
GWAS, integrative analysis, polygenicity, pleiotropy, Mendelian randomization

1 引言

人类基因组计划（human genome project，HGP）以及人类遗传变异图谱在21世纪初宣告完成。这一里程碑式的事件拉开了大规模利用生物医疗数据研究复杂人类疾病的序幕。随着生物科技（如基因芯片和测序技术）的飞速发展，全世界的研究者已经积累了多层面、多维度的基因大数据。这些数据覆盖从遗传变异到生命体表征的各个层面的数据，包括基因组（genome）层面、表观基因组（epigenome）层面、转录组（transcriptome）层面、蛋白质组（proteome）层面、代谢组（metabolome）层面以及生物体层面的表型特征（phenome），成为科学家研究复杂疾病的宝贵资源。无疑，基因大数据将对现有医学研究、个体化医疗产生颠覆性的影响。

全基因组关联分析（genome-wide association studies，GWAS）在基因大数据的研究中处于基础性地位。以GWAS为基础的大量研究项目的开展，开启了人类对各种复杂疾病的遗传结构的探索。GWAS采用高通量的方式获取全基因组的上百万个遗传变异位点——单核苷酸变异（single-nucleotide polymorphsim，SNP），并对这些变异点和复杂疾病或性状进行关联分析，找出风险变异点。截至2019年4月，已经公开发表的GWAS研究达3 923个，发现了约134 705个与疾病或性状显著关联的变异位点（显著性检验的p值$<5\times10^{-8}$）。大部分的研究结果可在全基因组关联分析数据库中获得。这些结果将基因组和生物体表型组联系起来，极大地增进了人们对生物体表型的遗传学结构的认识。随着GWAS研究越来越受到重视，大规模的研究也在不断推进。例如，英国生物样本库（UK Biobank）[1]成立于2006年，迄今为止，UK Biobank收集了约50万人的DNA测序数据以及大量与疾病和健康相关的表征数据，包括生理指标、生活方式、血液和尿液采样、大脑和身体成像数据以及参与者的医疗记录、电子病历数据等。这种从基因测序数据、表征数据到医疗成像数据和电子病历数据的多维度积累为科学家进行复杂疾病的研究提供了数据支撑，将助力于个性化的精准医疗服务。

尽管GWAS的研究取得了令人瞩目的成绩，然而单独的GWAS研究往往只考虑基因组层面与表型特征的信息，忽略了生命过程中其他层面的相关信息，无法贯穿从遗传变异到表型特征的整条证据链。因此需要整合和集成从遗传变异到生命体表征的多组学层面的信息并进行分析，从而全方位解析从遗传变异到表型特征的整个因果链条。这里的多组学层面包括基因组层面、表观基因组层面、转录组层面、蛋白质组层面、代谢组层面以及生命体表征组层面的表型特征（如**图1**所示）。伴随着各项大型生物医疗国际项目的开展，这些多层面、多维度、高质量的基因大数据正在不断产生和积累。2012年9月完成的DNA元件百科全书项目（encyclopedia of DNA elements project，ENCODE项目）[2]是继人类基因组计划后又一重要的突破性工程。该项目对大部分非编码序列（约占全基因组的98%）的功能进行了注释，例如“这部分序列与一种蛋白质结合”“这部分序列常被甲基基团标记”“这部分序列通常隐藏在组蛋白的包围中”。随之开展的表观基因组的图谱计划（roadmap epigenomics project）直接从人体的细胞系或组织中取样，并将其数据向公众开放。这些结果将为研究人类细胞系和组织的表观基因组功能发挥重要的作

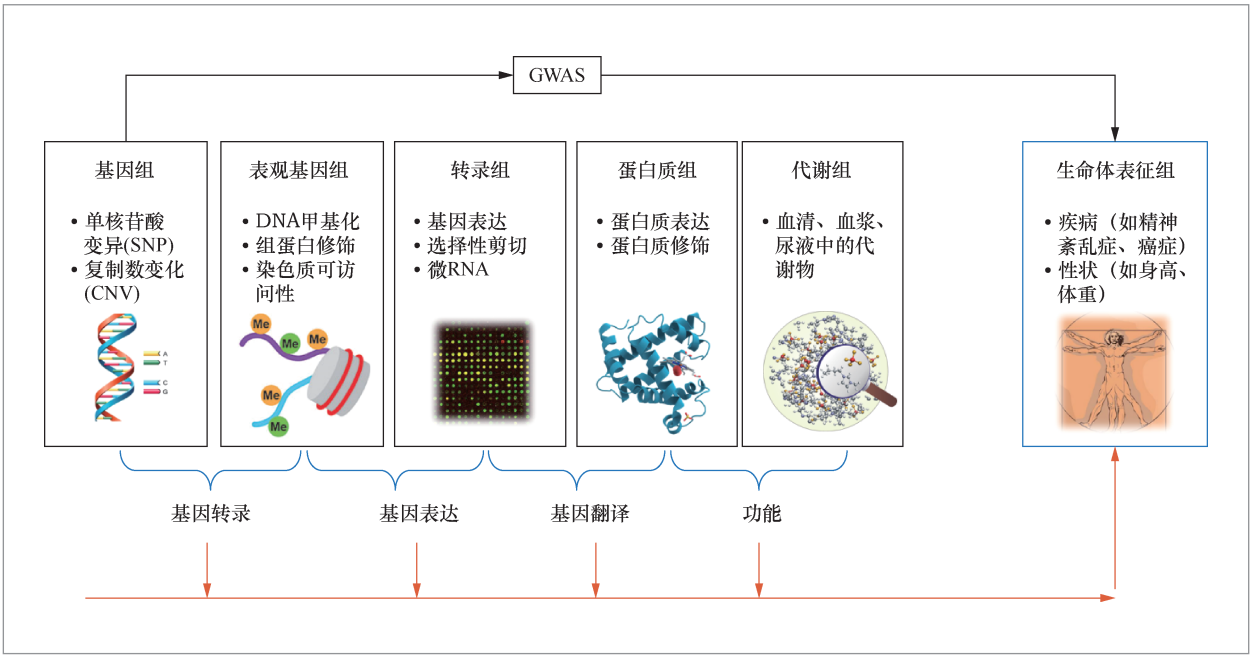


图1 多组学层面的数据

用。与此同时，2010年开始的基因型-组织表达（genotype-tissue expression，GTEX）项目[3]于2013公开数据库，到目前为止收集了约714个捐献者53个人体组织的11 688个样本数据，用于研究不同组织中基因型与表达型的关系，完成了基因组到转录组的跨越，为科学家提供了宝贵的资源库。此外，现代核磁共振技术、质谱和色谱等技术的发展使得对大规模的生物体小分子的定量研究成为可能，这也大大促进了蛋白质组层面和代谢组层面研究的发展。目前已公布的蛋白质层面和代谢组层面的研究达80多项，产生了大量可公开获取的数据。

这些多层面、多维度、高质量的基因大数据为刻画完整的疾病产生的因果链条创造了条件，同时也促进了该领域集成分析方法研究的发展。本文总结了当前基因大数据的集成分析的一些研究进展，这些研究大部分基于多基因遗传结构（多个基因影响一个性状）的共识以及以下两方面的科学发现：遗传变异的多效性（一个变异影响多种表型）以及非编码遗传变异的调控功能。具体地，本文对基因大数据的集成分析从以下3个方面进行综述：检测风险位点及其功能分析；基因多效性的分析；基于孟德尔随机化的因果推断。本文进一步结合具体的应用案例进行分析，最后对基因大数据的集成分析研究进行了总结以及展望。

2 基因大数据集成分析的科学依据

2.1 多基因遗传结构

复杂性状或疾病的多基因遗传结构是指复杂性状和疾病的变异是由多个基因共同影响的。尽管科学家在GWAS研究中发现了很多显著的变异位点，但是这些显著的变异位点只能解释性状方差的小部分，这个现象通常被称为“丢失的遗传

率（missing heritability）”。以人体身高这一性状为例，根据以往对家庭谱系的研究，已经存在的广泛认识是人类身高的遗传率达70%~80%[4]。2008年，一项关于身高的GWAS的研究发现了影响身高的40个位点，然而，这部分位点可以解释的遗传率只占5%[5]。到2014年，发现的关联位点达到700个，但依然只能解释约20%的遗传率[6]。这些结果表明，大量的变异位点同时作用于身高，而单个变异位点的贡献是极小的，身高这一性状具备明显的多基因遗传结构（如**图2**所示）。这种多基因遗传结构也在越来越多的复杂疾病的GWAS研究中得到证实，如代谢综合征[7]、精神疾病[8-9]等。多基因遗传结构的存在使得GWAS研究面临着诸多挑战：第一，多基因结构意味着单个基因对性状的贡献是微弱的，受GWAS样本量的限制，单个GWAS研究很难找到遗传变异中微弱的信号点；第二，常用的基于稀疏和强信号假设的建模方法在这里不再适用。**图2**中横轴为染色体编号，每个SNP按染色体上的位置排列，纵轴为对应SNP的GWAS统计检验的-lg(p)。图中顶部的点对应-lg(p)>30的SNP。

基于多基因遗传结构的假设，线性混合模型（LMM）开始受到关注[10-12]。以遗传率的估算为例，假设已经获取n个个体的M个SNP的标准化后的基因测序数据$G \in R^{n\times M}$，对应的表型（如身高）数据$y \in R^{n\times 1}$，$X \in R^{n\times p}$表示p个协变量信息（如年龄、性别和一些用以控制群体分层的主成分数据），它们之间的关系可以由线性混合模型来建立，具体如下：

$$y = X\cdot\beta + G\cdot u + e \qquad (1)$$

其中，$\beta \in R^{p\times 1}$是固定效应，$u \in R^{M\times 1}$是随机效应，e是由环境因素引起的随机误差。这里假设u和e都是正态分布的，即$u \sim \mathrm{N}\left(0, \frac{1}{M}\sigma_u^2 \mathrm{I}\right)$和$e \sim \mathrm{N}\left(0, \sigma_e^2 \mathrm{I}\right)$。式（1）建立起了基因型与表型的关系。由遗传率的定义（基因型方差在表型方差中的占比）及以上线性混合模型，可以得到遗传率的计算式为：

$$h^2 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \qquad (2)$$

限制极大似然估计（REML）和最大期望算法（expectation-maximization algorithm，EM）常被用于估算遗传率，如全基因组复杂性状分析（genome-wide complex trait analysis，GCTA）工具[9]、全基因组高效混合模型关联分析（genome-

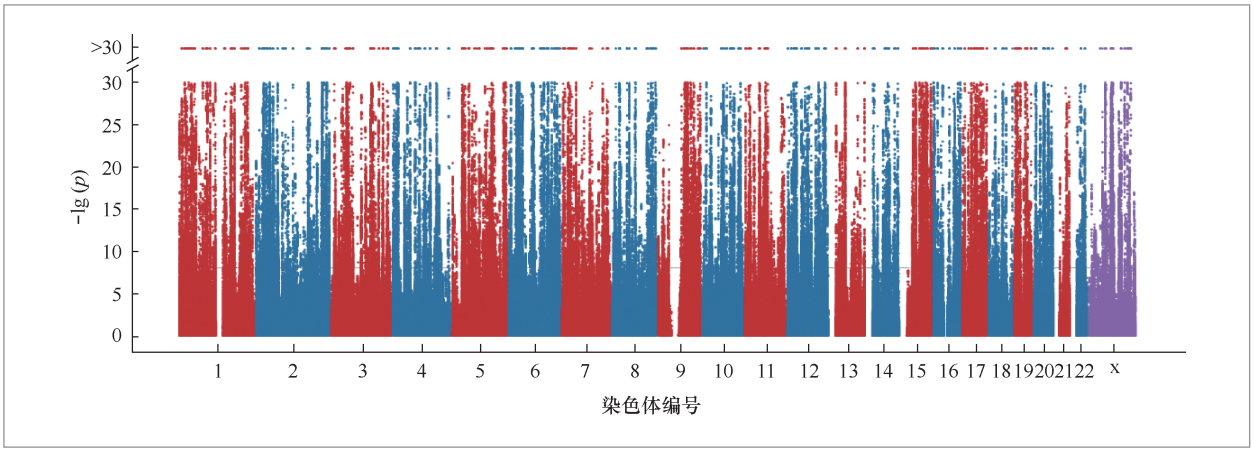


图 2 身高的曼哈顿图

wide efficient mixed model association, GEMMA)方法[11]。此外，LMM在GWAS研究中具备非常广泛的用途，如处理关联分析中的种群分层[11,13]、罕见变异的关联分析问题[14]，逐渐成为GWAS研究中集成分析的重要工具[15-16]。

2.2 基因的多效性

基因的多效性(pleiotropy)是指一个基因影响着多种表型。基因的多效性广泛存在于复杂性状中[17]，具体的例子有：与维生素D缺乏症强相关的CYP2R1基因会同时导致多发性硬化症[18-19]；CLPTM1基因同时影响神经胶质瘤、膀胱癌和肺癌[20]；基因多效性大量存在于自身免疫综合征和精神疾病中[21]。根据遗传变异影响性状的机制，基因多效性主要可分为两大类：一类是基因的生物多效性(biological pleiotropy)，即基因直接影响多种表型；另一类是基因的中介多效性(mediated pleiotropy)，即基因通过影响一种表型对另一种表型产生影响(如**图3**所示)。以肥胖基因(FTO基因)为例，研究者发现，该基因对骨关节炎的影响是通过影响体质指数(body mass index, BMI)来传递的[22]。基因的中介多效性在探索复杂疾病或性状之间的因果关系中有重要的作用[23]。

基因多效性的存在引导人们集成多个GWAS数据一起研究。一方面，基因的多效性可以帮助人们探索复杂疾病之间的关系，如遗传相关性、基于孟德尔随机化的复杂疾病的因果推断等；另一方面，由于多基因遗传结构的存在，单独分析一个GWAS会导致具备微弱效应的遗传位点很难被检测到。集成多个与遗传相关的复杂性状的GWAS数据，通过建立有效的统计模型，可以帮助检测到更多的多效性位点，从而提高统计分析的效率。

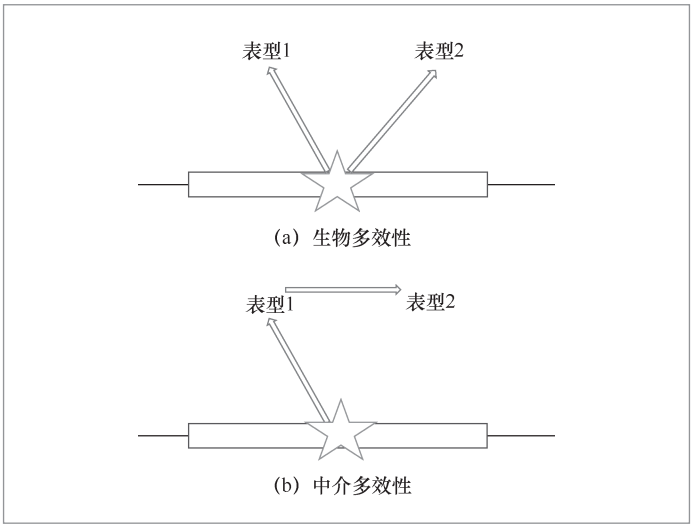


图 3 基因多效性

2.3 基因的调控功能

作为基本的遗传单位，基因是一段有功能性的DNA序列。基因中大部分位点(如启动子和增强子)处于非编码区，不参与基因编码，但对基因的转录、翻译和表达起着重要的调控作用。对于整体的DNA序列，只有少部分位于基因的调控区，参与基因编码，能够转录为信使RNA，进而指导蛋白质的合成。ENCODE项目对人类基因序列进行了详细分析，研究结果显示，在98%人类非编码区的序列中，81%保持着生物活性。大量证据显示，在GWAS发现的与众多复杂性状相关的位点中，大部分位点处在基因的非编码区。比如，科学家调查5 654个非编码区位点(这些位点与654种复杂疾病相关联)，发现76.6%的位点处于脱氧核糖核酸酶I高敏区(DHS)，这一区域与基因的转录功能高度相关[24]。基于转录组层面的研究发现，大量与疾病关联的位点位于表达数量性状基因座(expression quantitative trait loci, eQTL)(如**图4**所示)以及转录因子结合位点(transcription factor binding sites,

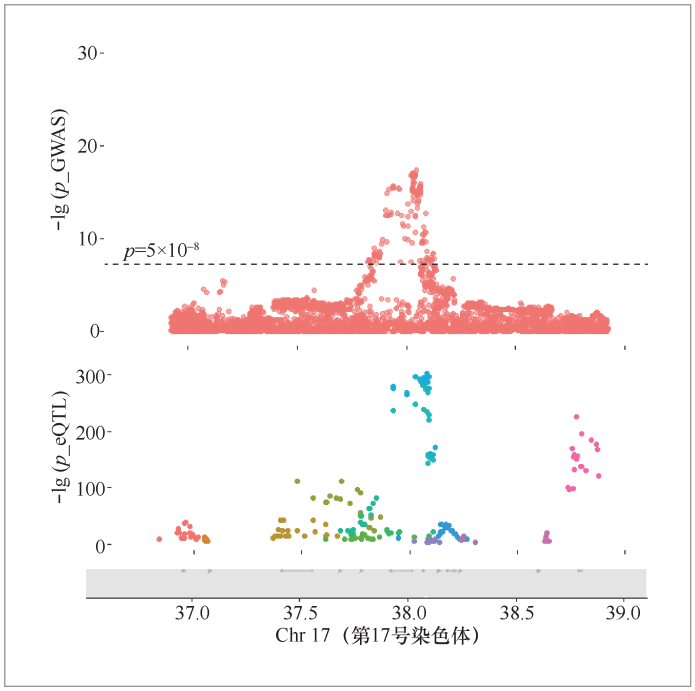


图 4 与疾病关联的位点常富集于表达数量性状基因座

TF）[25]。在最新的一项关于精神分裂症（schizophrenia，SCZ）的研究中发现，在与SCZ相关的100多个基因位点中，有超过20%的位点与基因的表达和调控机制相关[26]。这些研究结果表明，与疾病关联的风险变异点可能不直接参与基因编码，而是通过调控基因的转录、表达或翻译中的任何一个环节发挥作用。因此将疾病的GWAS数据和功能型数据集成并进行分析，能够深化人们对疾病的遗传机理的认识。

3 基因大数据的集成分析方法

基于上述科学进展，本节将从以下3个方面对基因大数据的集成分析进行综述。

- 检测风险位点及其功能分析：该类型方法集成GWAS数据和多组学数据（如变异位点功能型数据、转录组的基因表达数据等），力图找到更多的危险变异点，并对变异点的生理功能进行分析。
- 基因多效性的分析：该类方法集成多个GWAS数据以及多组学数据，从而探索复杂疾病遗传上的相关关系以及共同的致病基因。
- 基于孟德尔随机化的因果推断：该类型方法集成多种风险因素和疾病的GWAS数据，探讨风险因素（如血压、体重、代谢物等）对疾病的影响。

为描述方便，先对使用的数据类型进行介绍。在基因大数据的集成分析中使用的数据主要包括两种类型：第一类是个体样本层面的数据，该类型数据包含每个样本的基因型、每个样本的疾病状态（糖尿病、高血压或者健康）等；第二类是概括型数据，该类型数据是指对于个体样本层面的数据通过关联分析得到每一个SNP与疾病的概括性统计量，包括回归效应的估计值、标准差、检验统计量、p值等。因为使用的数据类型不同，集成分析方法也会存在差异。

3.1 风险遗传变异的检测及功能

由于与疾病相关的位点总是具备某些调控功能，将相关位点的调控功能信息植入某种疾病的GWAS数据分析中，能帮助人们找到真正的风险遗传变异点，深入地研究遗传变异的功能。在这个研究领域里，贝叶斯方法被大量运用到GWAS数据与调控信息数据的集成分析中[27]。该类方法对疾病的GWAS个体层面数据或概括性数据（用G表示）进行分析，同时引入功能型数据或基因表达数据（用A表示），推断遗传变异位点L（可以是一个基因也可以是单个SNP）是否为疾病发生的风险位点（见计算式（3））。基于模型分析的结果可以进一步对变异点进行功能分析。在这里，笔者介绍两类风险遗传变异检测的集成分析方法：第一类是结合功能型数据的

集成分析；第二类是结合基因表达数据的集成分析。

P(L是危险变异点 | GWAS数据G，功能型数据A) （3）

3.1.1 结合功能型数据的集成分析

假设已经收集了M个位点的D列功能型数据，并存放在矩阵A中，A中对应的第i行、第j列的元素记为A_{ij}，代表第i个SNP的第j种功能型的取值。以eQTL数据为例，假设A中的每一列对应人体的不同组织器官（如心脏、大脑、肝脏等）。如果第i个SNP能够调节第j个组织中某个基因的表达，那么这个SNP是第j个组织的一个eQTL，则A_{ij}取1，否则取0。将需要分析的SNP与已有的功能型数据库进行匹配，就可以产生矩阵A。下面以一种常用的统计模型[28]为例来阐述GWAS数据与功能型数据的集成分析。

假设观测到n个个体样本的表型数据（记为$y \in R^n$）及对应的基因型数据（用矩阵$X^{n\times M}$表示），那么，对于第i个样本个体，其表型数据y_i与其基因型数据X_{i1}，…，X_{iM}可以建立以下线性关系：

$$y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_M X_{iM} + \varepsilon_i \quad (4)$$

其中，$\beta_1,\cdots,\beta_M$为回归系数，ε_i为残差项，且服从正态分布$\mathrm{N}(0,\sigma^2)$。若回归系数β_j不为0，则代表第i个变异位点与疾病相关，因此进一步假设存在一个指示变量γ_j，γ_j可取0和1，分别代表β_j等于0和β_j不等于0。具体来说，β_j与γ_j的关系可以通过以下模型建立：

$$\begin{cases} \beta_j \sim \mathrm{N}\left(0,\sigma^2\right), \gamma_j=1 \\ \beta_j = 0, \qquad\quad \gamma_j=0 \end{cases} \quad (5)$$

接下来通过一个Logistic模型建立γ_j与第i个SNP的功能型数据信息A_j（即矩阵A中的第j行）的关系：

$$\log\frac{\Pr(\gamma_j=1\mid A_j)}{\Pr(\gamma_j=0\mid A_j)} = A_j\theta + \theta_0 \quad (6)$$

其中，$\theta \in R^K$和θ_0是需要估计的参数。若θ中的某个元素不为0，代表疾病的致病基因可能与对应的功能组相关。也就是说，如果知道某个SNP的功能信息，引入这些信息能够调整γ_j的先验分布，模型会自动给与疾病相关性高的某种功能型的位点赋予更高的权重，从而提高统计分析的效率。若模型的参数估计结果已经得到，通过计算在给定y、G、A下γ_j取1或0的后验概率，可以推断γ_j取1或0的概率，从而可以检测该变异位点是否属于风险变异点。对模型的参数进行统计推断可以对该变异位点进行功能分析。

上述模型的框架可以扩展到GWAS概括性数据建模分析中，如可以对概括性统计量的假设检验的p值进行建模，通过假设p值来自一个由均匀分布和Beta分布组成的混合分布[29]，建立起p值（如p_i）与指示变量γ_i的联系。具体来说，假设观测得到的p值用p_i,…，p_n表示，即与疾病不相关的SNP来自[0,1]区间的均匀分布U(0,1)，对应的指示变量γ_j此时取1，而与疾病相关的SNP的p值来自Beta分布$\mathrm{Beta}(a,1)$，对应的指示变量取0，具体如下：

$$\begin{cases} p_j \sim \mathrm{Beta}(a,1), & \gamma_j=1 \\ p_j \sim \mathrm{U}(0,1), & \gamma_j=0 \end{cases} \quad (7)$$

类似地，可以通过Logistic模型（见计算式（6））建立起γ_j与第j个SNP的功能型数据信息A_j的联系。最大期望算法常用于求解上述模型的参数，目前的研究已经可以满足集成大规模功能型数据的分析和计算[29]的需求。

近年来，深度学习作为一种机器学习技术开始得到越来越多的关注，在图像识别和自然语言处理等领域取得了突破性进展。因其优异的数据处理能力，深度学习也开始被逐步应用于探索人类基因组密码[30-31]，尤其是功能基因（functional genomics）组学。例如，2015年提出的DeepSEA[32]，该方法将公开的多组学功

能型数据（ENCODE功能型数据，表观基因组的图谱计划）作为输入，搭建深度学习模型，从而实现对非编码区域遗传变异位点（占人体基因组的98%）的功能分析。2016年提出的DeepWAS[33]首先基于DeepSEA找到潜在的影响疾病的某一类型的功能型位点，然后再构造带惩罚函数的多元回归模型（LASSO），实现风险变异位点的检测。单独的GWAS分析只能确定某个变异点与疾病的关联性，DeepWAS则能够更加综合地分析致病基因突变，帮助寻找其致病机理。实际应用显示，集成分析GWAS数据和功能型数据不仅可以帮助检测新的危险变异，而且对探索疾病的发病机制有非常重要的作用[21,33]。

3.1.2 结合基因表达数据的集成分析

结合基因表达数据的集成分析方法有PrediXcan[34]、MetaXcan[35]、TWAS[36]以及CoMM[37]。该类型方法的基本思想是以已有的基因表达样本库的数据为参考面，估算GWAS的个体的基因表达，然后检测基因表达和表征是否关联，其基本原理如**图5**所示，此类方法的目的在于对基因调控的表达部分和性状进行关联分析（**图5**中实线箭头）。以CoMM为例[37]。假设参考面的某个基因的基因表达数据以及基因型数据为y_{1g}和X_{1g}，感兴趣的疾病或性状的表型和基因型数据为y_2和X_{2g}，这里基因的表达数据和GWAS个体层面的数据是不同的样本，基因表达型数据和GWAS数据的关系通过共享参数u来建立，具体如下：

$$\begin{aligned} y_{1g} &= X_{1g} \cdot u + e_1 \\ y_2 &= \alpha_g \cdot X_{2g} \cdot u + e_2 \end{aligned} \quad (8)$$

其中，α_g为未知参数，通过检验α_g是否为0，可以推断基因表示与表征的关系。基于多基因效应的假设，这里假设u服从以0为均值，以$\sigma_u^2 I$为方差的多元正态分布，即$u \sim \mathrm{N}(0, \sigma_u^2 I)$。$e_1$、$e_2$为误差项且服从零均值正态分布。该方法将基因表达数据和疾病或性状的GWAS数据进行联合建模和估计，考虑了估算基因表达的不确定性，显示出更高的统计效率。这类型方法也可以推广到其他层面的组学数据（如蛋白组层面）和GWAS的集成分析研究中，具有极高的应用价值。

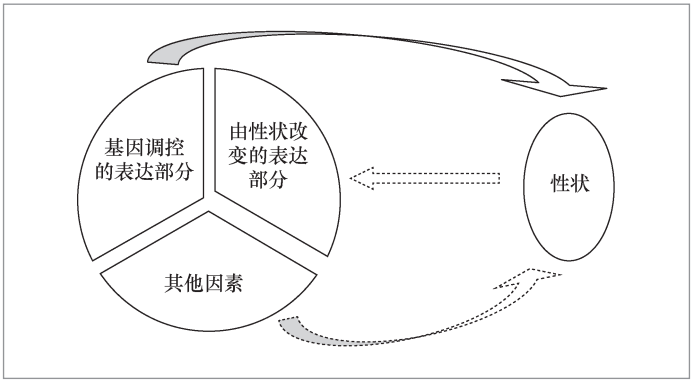


图 5 基因表达的数据和疾病的 GWAS 数据的集成分析基本原理

3.2 遗传变异的多效性

遗传变异的多效性的存在决定人们可以集成多个GWAS数据进行分析。一方面，可以探讨复杂疾病在遗传结构上的相关关系，这种相关关系在某种程度上可以反映复杂疾病之间的联系；另一方面，利用这种多效性设计有效的统计模型，可以帮助人们找到更多与多种复杂疾病相关的多效性位点，同时，通过引入功能型数据进行分析，能进一步提高人们对疾病的遗传机理的认识。

复杂疾病在遗传结构上的相关关系可以通过遗传相关系数这一参数来度量，该参数反映两个复杂性状中由遗传效应引起的总体相关性程度。常见疾病（如心脏病、糖尿病、脂类代谢异常以及高血压）两两之间存在较强的遗传相关系数（0.27~0.43），反映了遗传变异在这一类型的疾病中具有比较一致的效应。基于线性混合模型，可以先了解遗传相关系数的定

义。假设已经获取两个独立样本的关于两个表征的观测数据以及个体层面的基因测序数据（标准化后），表示为：$y_1 \in R^{n_1 \times 1}$，$G_1 \in R^{n_1 \times m}$，$y_2 \in R^{n_2 \times 1}$，$G_2 \in R^{n_2 \times 1}$，对应的协变量信息用$X_1 \in R^{n_1 \times p_1}$、$X_2 \in R^{n_2 \times p_2}$表示，这里的下标对应两个不同表征型，它们之间的关系由以下线性混合模型来建立：

$$\begin{aligned} y_1 &= X_1\beta_1 + G_1u_1 + e_1 \\ y_2 &= X_2\beta_2 + G_2u_2 + e_2 \end{aligned} \quad (9)$$

基于多基因效应，假设：

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right) \quad (10)$$

其中，σ_1^2、σ_2^2分别为u_1和u_2的方差，ρ为u_1和u_2的相关系数，即遗传相关系数，其取值范围为[-1,1]。极大似然估计或限制性极大似然估计等方法常用于二元混合模型的参数估计。此外，基于概括性数据的LDScore回归方法[38]，可以针对任意两个性状进行分析，并且能有效地处理样本重合的影响。在该方法的基础上，研究者们提出了GNOVA[16]，该方法可以计算不同性状在功能型位点的分区遗传相关程度，能够帮助人们更具体地了解疾病的遗传特征。

更重要的是，集成多个GWAS数据的分析方法可以帮助人们进行多效性位点的检测。研究者们通过对比不同性状或疾病的GWAS研究的显著位点，可以标注多效性基因位点。然而这种方法只考虑了显著位点，因而缺乏统计效率。通过集成分析多个不同性状的全基因组的GWAS数据（不只考虑显著位点），建立有效的统计模型，可以检测更多的风险变异点，并发现更多的多效性位点，进而提高统计分析的效率。多变量的线性混合模型也常被用于关联变异位点的检测，显示出比一元线性混合模型更大的统计效率。目前常用的工具有GEMMA[11]、mvLMM[39]和BOLT-REML[40]等。该类型方法在实际应用中往往受到很多限制，原因在于这类型方法需要用到GWAS个体层面的数据。受数据可获得性和数据隐私保护等的限制，获取GWAS个体层面的数据往往会比较困难。基于概括性数据的多个GWAS的集成分析，只需要GWAS概括性数据，而且不同性状的GWAS数据可以来自不同的生物样本集（考虑到基因的种群效应，一般要求GWAS数据来自同种群），有利于大规模的集成分析。同时，该类型方法能够更灵活地引入功能型数据，提高模型的统计效率。基于隐变量的统计模型在GWAS的集成分析中扮演着重要的角色。以基于基因多效性和功能型数据的遗传分析（genetic analysis incorporating pleiotropy and annotation, GPA）方法[41]为例，该方法基于概括性统计量（p值），集成分析多个GWAS数据和功能型数据，能有效地检测出多效性位点。假设收集到K个GWAS的M个SNP的p值，存放在一个$M \times K$维的矩阵中，用P表示，其中p_{jk}代表第j个SNP在第k个GWAS的p值。类似于单个GWAS危险变异的检测，这里也假设存在一个$M \times K$维的隐变量矩阵Z，Z_{jk}取0代表第j个SNP和第k个GWAS无关，Z_{jk}取1代表第j个SNP和第k个GWAS有关。变量之间对应的关系可以通过下列模型表示：

$$\begin{cases} P_{jk} \sim \mathrm{Beta}(a_k,1), & Z_{jk}=1 \\ P_{jk} \sim \mathrm{U}(0,1), & Z_{jk}=0 \end{cases} \quad (11)$$

以$k=2$为例，定义：

$$\begin{aligned} &\pi_{00} = \Pr(Z_{j1}=0, Z_{j2}=0) : P_{j1} \sim \mathrm{U}(0,1), \\ &P_{j2} \sim \mathrm{U}(0,1) \mid Z_{j1}=0, Z_{j2}=0 \\ &\pi_{01} = \Pr(Z_{j1}=0, Z_{j2}=1) : P_{j1} \sim \mathrm{U}(0,1), \\ &P_{j2} \sim \mathrm{Beta}(a_2,1) \mid Z_{j1}=0, Z_{j2}=1 \\ &\pi_{10} = \Pr(Z_{j1}=1, Z_{j2}=0) : P_{j1} \sim \mathrm{Beta}(a_1,1), \\ &P_{j2} \sim \mathrm{U}(0,1) \mid Z_{j1}=1, Z_{j2}=0 \\ &\pi_{11} = \Pr(Z_{j1}=1, Z_{j2}=1) : P_{j1} \sim \mathrm{Beta}(a_1,1), \\ &P_{j2} \sim \mathrm{Beta}(a_2,1) \mid Z_{j1}=1, Z_{j2}=1 \end{aligned} \quad (12)$$

这里π_{00}、π_{01}、π_{10}、π_{11}分别表示Z_{j1}、Z_{j2}取不同值的比例，并且$\pi_{00}+\pi_{01}+\pi_{10}+\pi_{11}=1$，进一步地，GPA假设在给定GWAS的关联状态下，SNP的功能性状态之间独立。

$$\begin{aligned} q_{00k} &= \Pr(A_{jk}=1 \mid Z_{j1}=0, Z_{j2}=0) \\ q_{01k} &= \Pr(A_{jk}=1 \mid Z_{j1}=0, Z_{j2}=1) \\ q_{10k} &= \Pr(A_{jk}=1 \mid Z_{j1}=1, Z_{j2}=0) \\ q_{11k} &= \Pr(A_{jk}=1 \mid Z_{j1}=1, Z_{j2}=1) \end{aligned} \quad (13)$$

在GPA模型中，π_{00}、π_{01}、π_{10}、π_{11}、q_{00k}、q_{01k}、q_{10k}、q_{11k}以及a_1、a_2为模型的未知参数，EM类型的算法可以有效地解决该模型的参数估计问题。模型通过计算局部错误发现率（false positive rate，FDR）$\Pr(Z_{j1}, Z_{j2} \mid P, A)$、$\Pr(Z_{j1} \mid P, A)$和$\Pr(Z_{j2} \mid P, A)$来检测多效性位点以及单个GWAS的风险变异点；对$q_{00k}$、$q_{01k}$、$q_{10k}$、$q_{11k}$进行显著性检验可以检测与疾病相关联的位点在第$k$类功能型上的富集情况。

进一步的研究显示，通过Logistic回归[42]和Probit[43]回归实现隐变量和功能型数据的建模，可以解决GPA模型只能引入离散型、功能型数据的局限。同时，在多个GWAS数据的集成分析中，低秩性和稀疏性的叠加模型以及复合极大似然方法的使用，完美地避免了模型参数指数增长而无法估计的问题。

3.3 基于孟德尔随机化法的复杂性状因果推断

基因的多效性普遍存在于人类基因组中，对复杂疾病的研究具有重要的价值。基因的中介多效性可以用来探索复杂疾病或性状之间的因果关系，这类型的方法被称作孟德尔随机化法（Mendelian randomization，MR）。该方法可追溯至1986年，Katan M B[44]提出载脂蛋白E（APOE）基因的变异（包括E2/E3/E4）能够影响体内胆固醇的水平，其中携带E2基因的人血浆胆固醇水平较低。由于亲代到子代等位基因的随机分配，个体APOE基因的携带情况不受其他混杂因素的影响。因此，可通过癌症病人和非癌症病人的E2基因携带情况是否存在差异来判断低水平血浆胆固醇是否会增加癌症风险。孟德尔随机化法这一术语在1991年第一次被提出[45]，逐渐成为因果推断中的主流方法[46]。该方法以基因变异（如APOE）为工具变量来研究暴露因素（如血浆胆固醇水平）和结局变量（如癌症）之间的因果关系，可以突破传统的观察性流行病学在研究因果关系方面的多种限制，如反向因果、混杂因素等。近几年来，基于GWAS概括性数据的MR分析方法的发展更能体现出其方法上的优越性。首先，大量GWAS概括性数据的可公开获得为研究者们提供了丰富的数据资源。MR检验可以在任意的性状之间进行，大大扩展了因果关系的可研究范围。相比之下，传统的随机控制实验往往受到实验伦理以及试验设计的局限，对于罕见疾病的研究，往往因数据收集方面的困难而无法进行；其次，GWAS的研究基本上是基于大样本的研究，目前还在不断朝着更大样本规模发展，MR因果分析也将受益于次。

孟德尔随机化法以遗传变异（G）为工具变量，在推断暴露因素（X）对结局变量（Y）的因果关系时，需要满足以下3条基本条件：①G和X相关；②假设存在混杂因素（U）同时影响暴露因素和结局变量，G和U无关；③G只能通过X影响Y，而不能存在其他的通道影响Y。只有在这3个条件同时满足的情况下，MR方法才能正确地推断X和Y的因果关系（如图6所示）。具体来说，条件①要求在做MR分析时，必须选择与X显著相关的位点，比如$p<5\times10^{-8}$。根据孟德尔遗传定律，总是假设条件②成立，然而其他因素（如群体分层、样本的重合）可能会导致这个条件不符合，因此在实际应

用中选择的GWAS数据需来自同一种群且应避免存在样本重合；条件③要求G对Y不能有直接的影响，G对X和Y只能存在中介多效性。基因中介多效性也常被称作垂直多效性，对应的基因的生物多效性也称作水平多效性，用以描述基因与性状之间的直接联系。水平多效性的存在会影响MR分析结果的可靠性[47]。

MR分析中各变量间（包括G、X、Y、U）的关系可通过下面的线性结构方程来描述：

$$\begin{aligned} X &= \sum_{j=1}^{n}\gamma_j G_j + \eta_x U + E_x \\ Y &= \beta X + \sum_{j=1}^{n}\alpha_j G_j + \eta_y U + E_y \end{aligned} \quad (14)$$

其中，$\{G_j\}_{j=1}^{M}$为M个独立的变异位点，γ_j为G_j对X的效应大小（effect size），α_j为G_j对Y的效应大小。若该变异位点满足MR分析的假设，则α_j=0；若存在水平多效性，则$\alpha_j \neq 0$。η_x和η_j表示混杂因素对X和Y的影响大小，E_x和E_y为对应的残差向量，β表示X对Y的影响效应的大小，若β不为0，则表示X和Y存在因果关系。MR分析的目的是准确地估计β的大小并进行因果关系的检验。上述模型要求GWAS个体层面的数据，而基于两样本的GWAS概括性统计量的MR分析方法不受数据的隐私保护等限制，逐渐成为热点。这时，观测到的数据为变异位点对暴露因素（X）和结局变量（Y）的回归效应（effect）的估计值及其标准差，记为$\{\hat{\gamma}_j, \sigma_{Xj}\}_{j=1}^{M}$和$\{\hat{\Gamma}_j, \sigma_{Yj}\}_{j=1}^{M}$，它们的关系可以表示为：

$$\begin{aligned} \hat{\gamma}_j &\sim \mathrm{N}\left(\gamma_j, \sigma_{Xj}\right) \\ \hat{\Gamma}_j &\sim \mathrm{N}\left(\beta\gamma_j + \alpha_j, \sigma_{Yj}\right) \end{aligned} \quad (15)$$

由于两组数据来自无样本重合的GWAS研究，$\hat{\gamma}_j$与$\hat{\Gamma}_j$相互独立（给定的情况下）。若变异位点均满足MR假设条件，不存在水平多效性，则$\alpha_j = 0,\ j = 1, \cdots n$；若存在某几个$\alpha_j$不为0，由于会影响$\beta$的估计，则会被当作异常点。若所有位点均存在水平多效性但平均水平为0，则称为平衡的水平多效性或系统性多效性，反之则称为非平衡的水平多效性或有向多效性[48-49]。

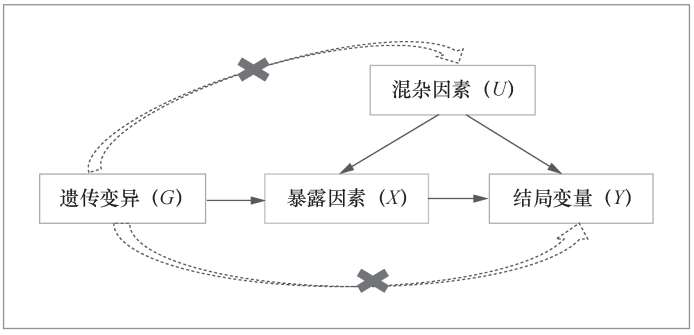


图6 孟德尔随机化法基本原理

目前关于MR的大量研究集中在处理变异位点的水平多效性方面，研究方法可大致分为3种类型。一类是矫正，如MR-Egger[49]通过在传统的异方差加权模型（inverse variance weighted，IVW）中引入常数项，修正由于非平衡的水平多效性（α_j的均值不为0）对模型带来的影响。该方法往往会引入较大的估计方差，降低统计效率。第二类是异常值剔除，这类型方法（如MR-PRESSO[50]、GSMR[51]等）先对可能存在水平效应的异常位点进行检测，剔除检测到的异常位点，然后对模型参数重新估计。这类方法在存在少量的异常点时是有效的，但当存在大量的异常点或平衡的水平效应时，则需谨慎使用。第三类方法是对违背MR假设的变异点进行具体的建模，如MR.raps[52]和BWMR[53]等。MR.raps和BWMR均假设存在系统性的水平效应，而且α_j服从零均值的正态分布，MR.raps在使用似然方法进行参数估计时，用稳健的损失函数Tukey损失函数和Huber损失函数取缔了均方损失函数，以得到稳健的估计。BWMR则引入贝叶斯重加权

（Bayesian reweighting）的思想，通过赋予异常点较低的权重来降低异常点的影响。

基于概括型统计量的MR分析，集成来自不同样本的GWAS研究数据，可以推断出不同性状之间的因果关系，对了解复杂疾病的发病机制、疾病的干预治疗以及制药等方面具有重要的推进作用。

4 应用案例

4.1 基因多效性分析：运用于神经质和重度抑郁症GWAS数据分析

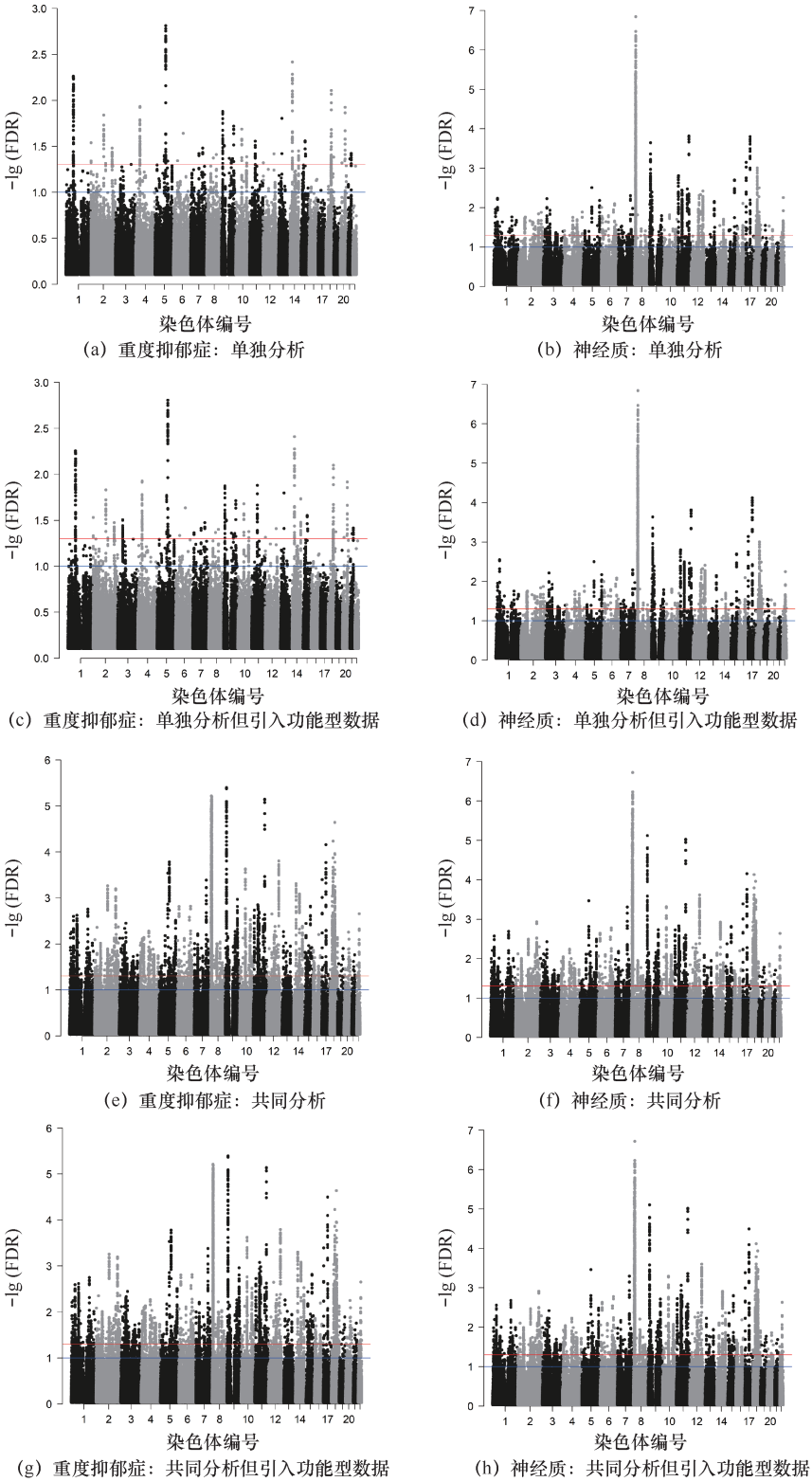
心理学上描述人的五大性格特质包括：开放性、责任性、外倾性、宜人性和神经质，其中神经质主要反映平衡焦虑、敌对、压抑、自我意识、冲动、脆弱等情绪的能力，即维持情绪稳定性的能力。已有的研究表明，神经质[54]和重度抑郁症（major depression disorder，MDD）[55]存在较强的关联[56]，在遗传上存在较强的相关性[57]。本文基于GPA[41]方法对已有的神经质和重度抑郁症的GWAS概括型数据和功能型数据（这里使用eQTL数据）进行基因的多效性分析，以探索这两种性状在遗传上的关系。集成分析结果见**表1**，曼哈顿图如**图7**所示。集成两组GWAS数据进行分析显示出更高的统计效率，原因在于这两组疾病具有高度相似的遗传学机理。根据参数估计结果，$\hat{\pi}_{00}=0.793$，$\pi_{01}=0.014$，$\pi_{10}=0$，$\pi_{11}=0.194$，其中，$\hat{\pi}_{11}=0.194$（似然比检验$p=0$）表明神经质和重度抑郁症具有高度相似的遗传机理，也就是说两个性状共享信息，共同分析两个GWAS数据能极大地提高效率。此外，**表1**的结果显示通过集成功能型数据，能发现更多的显著位点，进一步提高统计效率。

表1 神经质和重度抑郁症的集成分析结果

疾病	与另一性状同时分析	引入功能型数据	显著的变异位点数（FDR< 0.05）/个
神经质	否	否	1 623
重度抑郁症	否	否	322
神经质	否	是	1 653
重度抑郁症	否	是	353
神经质	是	否	3 207
重度抑郁症	是	否	3 739
神经质	是	是	3 281
重度抑郁症	是	是	3 818

4.2 血脂代谢物与人类复杂疾病的因果关系

本节基于101种分子水平的血脂代谢物[58]和57种人类复杂疾病的GWAS概括型数据，探讨集成分析在实际中的应用。复杂性状或疾病的GWAS数据来源见**表2**，血脂代谢物包含87种血脂蛋白携带的脂类物质以及14种游离的脂肪酸或相关度量。脂类物质包括总胆固醇（total cholesterol，C）、胆固醇酯（cholesterol esters，CE）、总脂质（total lipids，L）、磷脂（phospholipids，P）、游离胆固醇（free cholesterol，FC）、甘油三酯（triglycerides，TG）。血脂蛋白根据密度可分为低密度脂蛋白（low-density lipoprotein，LDL）、高密度脂蛋白（high-density lipoprotein，HDL）、极低密度脂蛋白（very-low lipoprotein，VLDL）、中密度脂蛋白（intermediate-density lipoprotein，IDL）。各类血脂蛋白根据大小可以进行细分，如LDL可分为L.LDL、M.LDL、S.LDL等。57种人类复杂疾病包括与人体测量相关的性状（如身高、体脂指数（BMI）等）、与心血管相关性状（如冠心病（CAD）、收缩压（SBP）、舒张压



（a）重度抑郁症：单独分析　（b）神经质：单独分析

（c）重度抑郁症：单独分析但引入功能型数据　（d）神经质：单独分析但引入功能型数据

（e）重度抑郁症：共同分析　（f）神经质：共同分析

（g）重度抑郁症：共同分析但引入功能型数据　（h）神经质：共同分析但引入功能型数据

图 7　神经质和重度抑郁症的曼哈顿图（蓝线和红线分别对应局部错误率为 0.1 和 0.05）

表 2 复杂性状和疾病的 GWAS 数据来源

类别	性状或疾病	中文名称	英文名称	参考文献
人体测量指标	BMI	体重指数	body mass index	[59]
	BFP	体脂百分比	body fat percentage	[60]
	Hip	臀围	hip circumference	[61]
	WC	腰围	waist circumference	[61]
	WHR	腰臀比	waist hip ratio	[61]
	BL	出生长度	birth length	[62]
	BW	出生体重	birth weight	[63]
	CO	童年肥胖	childhood obesity	[64]
	IHC	婴儿头围	infant head circumference	[65]
	Height	身高	height	[66]
与心血管相关的指标	DBP	舒张压	diastolic blood pressure	[67]
	HTN	高血压	hypertension	[67]
	MAP	平均动脉压	mean arterial pressure	[67]
	PP	脉压	pulse pressure	[67]
	SBP	收缩压	systolic blood pressure	[67]
	PVD	周边血管疾病	peripheral vascular disease	[51]
	CAD	冠心病	coronary artery disease	[68]
	HR	心率	heart rate	[69]
	HRV_Pvrsa_Hf	心率变异性指标一	heart rate variability index 1	[70]
	HRV_Rmssd	心率变异性指标二	heart rate variability index 2	[70]
	HR_Sdnn	心率变异性指标三	heart rate variability index 3	[70]
免疫系统疾病	AD	特应性皮炎	atopic dermatitis	[71]
	Crohn_Disease	克罗恩病	Crohn disease	[72]
	Celiac_Disease	腹腔疾病	celiac disease	[73]
	Eczema	湿疹	eczema	[71]
	IBD	炎症性肠病	inflammatory bowel diseases	[72]
	MS	多发性硬化症	multiple sclerosis	[74]
	PBC	原发性胆汁性肝硬化	primary biliary cirrhosis	[75]
	RA	类风湿关节炎	rheumatoid arthritis	[76]
	UC	溃疡性结肠炎	ulcerative colitis	[72]
	T1D	Ⅰ型糖尿病	type 1 diabetes	[77]
	SLE	系统性红斑狼疮	systemic lupus erythematosus	[78]
代谢类疾病	T2D	Ⅱ型糖尿病	type2 diabetes	[79]
	ANM	更年期年龄	age at natural menopause	[80]
	Dyslipid	脂类代谢异常	dyslipidemia	[51]
	eGFR	肾小球滤过率	estimated glomerular filtration rate	[81]
	FPI	空腹胰岛素原	fasting proinsulin	[82]
	uACR	尿蛋白/肌酐比值	urinary albumin to cretinine ratio	[83]
	AM	初潮年龄	age at menarche	[84]
	FI	空腹胰岛素	fasting insulin	[85]
神经系统疾病	AD	阿尔茨海默症	alzheimer's disease	[86]
	ALS	肌萎缩侧索硬化症	amyotrophic lateral sclerosis	[87]
	MD	黄斑变性	macular degeneration	[51]
	Parkinson	帕金森综合征	Parkinson's disease	[88]
精神疾病	Angst	焦虑症	angst	[89]
	BIP	双相情感障碍症	bipolar disorder	[90]
	ADHD	多动症	attention deficit hyperactivity disorder	[55]
	ASD	自闭症谱系障碍	autism spectrum disorder	[55]
	PTSD	创伤后应激障碍	post-traumatic stress disorder	[91]
	SCZ	精神分裂症	schizophrenia	[55]
	Stress	压力	stress	[51]
	Depress	抑郁	depress	[51]
	AN	神经性厌食症	anorexia nervosa	[92]
	CAB	童年期攻击行为	childhood aggressive behaviour	[93]
	Loneliness	孤独	loneliness	[94]
	OCD	强迫症	obsessive compulsive disorder	[95]
	MDD	重度抑郁症	major depressive disorder	[95]

（DBP）等）、代谢类疾病（如Ⅱ型糖尿病、脂类代谢异常等）、中枢神经系统疾病（如阿尔茨海默症、帕金森综合征等）、精神疾病（如重度抑郁症、多动症等）以及自身免疫疾病（如Ⅰ型糖尿病、系统性红斑狼疮等）。具体的分析分两方面：一方面，基于基因的多效性系统地探讨血脂代谢物和复杂疾病在遗传上的相关关系，对了解复杂疾病的遗传结构和性质具有重要的作用，本文基于GNOVA[16]方法，对血脂代谢物和与人类复杂疾病的遗传相关系数进行估计；另一方面，血脂代谢物和复杂疾病的因果关系一直是科学家们关注的重点，对于疾病的预防和诊断具有重要的指导作用。以心血管疾病为例，目前大量的研究表明，血液中的低密度胆固醇是造成心血管疾病的主要原因。本文应用BWMR[53]方法，对血脂代谢物和人类复杂疾病的因果关系和反向因果关系进行分析，从而深化对血脂代谢物和复杂疾病的认识。

图8展示的是通过Bonferroni校正之后显著的遗传相关系数（$p<0.05/(101\times57)$），其中640对具备显著的遗传相关关系。**图8**中红色为正相关，绿色为负相关，颜色深度和方格面积表示相关系数的大小，颜色越深或方格面积越大表示遗传相关系数越大。遗传相关系数的估计需调用GNOVA方法的软件包。结果显示，大部分复杂疾病与大部分血脂代谢物水平均有显著的遗传上的相关关系，如心血管疾病、高血压疾病、血压相关度量、体重相关度量等。此外，阿尔茨海默症、神经性厌食症与多种类型的血脂代谢物有密切的关系。值得注意的是，双相情感障碍症与多种游离脂肪酸有显著的遗传上的相关关系。因果分析结果显示，体质指数、Ⅱ型糖尿病会影响体内血脂代谢物的水平（结果如**图9**所示）。

5 结束语

本文对GWAS研究中的集成分析进行了综述，主要应用于以下3个方面：检测风险位点及其功能分析、基因多效性的分

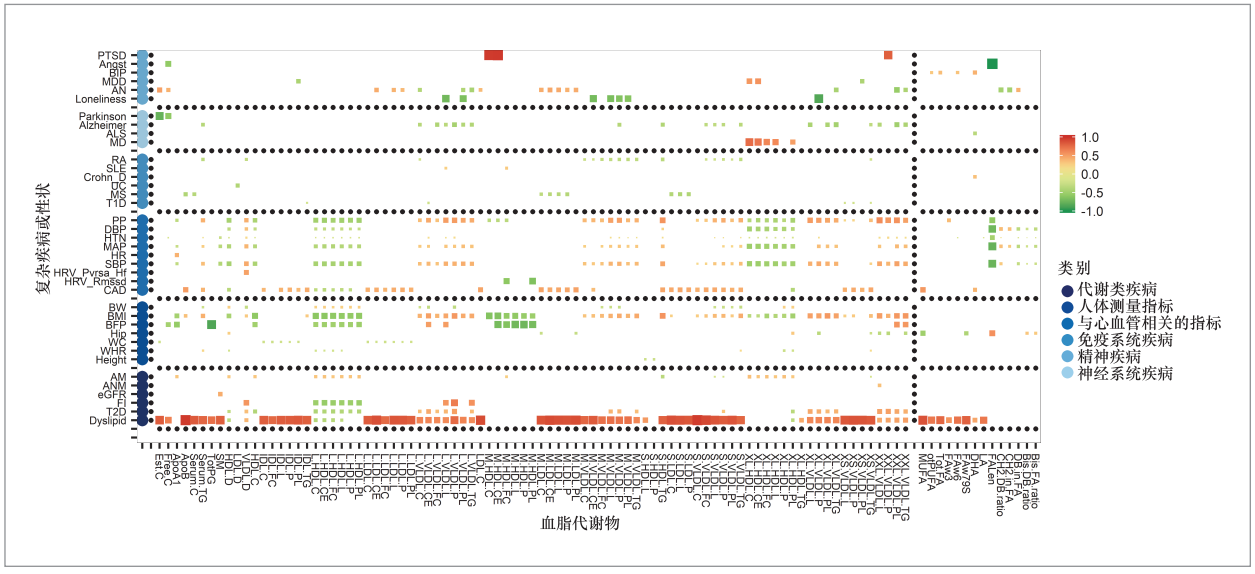


图 8　血脂代谢物（101 种）与复杂疾病或性状（57 种）的相关系数

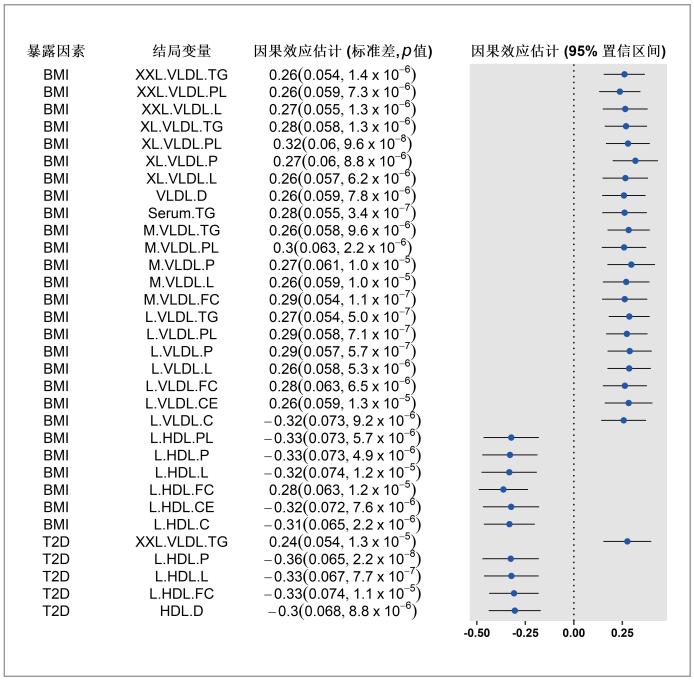


图 9 血脂代谢物和复杂性状的因果分析结果

析、基于孟德尔随机化的因果推断。实际应用显示，集成分析在GWAS研究中显示出重要的作用，有助于挖掘重要的信息。同时，多种类型的集成分析统计方法具备极大的拓展空间，将对未来的研究发挥更大的价值。未来GWAS的集成分析将更多地应用于集成多组学数据的分析，对探索从遗传变异到疾病发生的整个因果链条起着重要作用。随着越来越多的高质量数据的不断产生，全方位生物医疗大数据（包括基因大数据、医疗图像数据、电子病历等）的集成分析将使得实现个性化的精准医疗成为可能。

参考文献:

[1] SUDLOW C, GALLACHER J, ALLEN N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age[J]. PLoS Medicine, 2015, 12(3): e1001779.

[2] ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project[J]. Science, 2004, 306(5696): 636-640.

[3] LONSDALE J, THOMAS J, SALVATORE M, et al. The genotype-tissue expression (GTEx) project[J]. Nature Genetics, 2013, 45(6): 580.

[4] VISSCHER P M, MCEVOY B, YANG J. From Galton to GWAS: quantitative genetics of human height[J]. Genetics Research, 2010, 92(5-6): 371-379.

[5] MANOLIO T A, COLLINS F S, COX N J, et al. Finding the missing heritability of complex diseases[J]. Nature, 2009, 461(7265): 747.

[6] WOOD A R, ESKO T, YANG J, et al. Defining the role of common variation in the genomic and biological architecture of adult human height[J]. Nature Genetics, 2014, 46(11): 1173.

[7] VATTIKUTI S, GUO J, CHOW C C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits[J]. PLoS Genetics, 2012, 8(3): e1002637.

[8] LEE S H, DECANDIA T R, RIPKE S, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs[J]. Nature Genetics, 2012, 44(3): 247-250.

[9] LEE S H, RIPKE S, NEALE B M, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs[J]. Nature Genetics, 2013, 45(9): 984-994.

[10] YANG J, LEE S H, GODDARD M E, et al. GCTA: a tool for genome-wide complex trait analysis[J]. The American Journal of Human Genetics, 2011, 88(1): 76-82.

[11] ZHOU X, STEPHENS M. Genome-wide efficient mixed-model analysis

for association studies[J]. Nature Genetics, 2012, 44(7): 821.

[12] LEE S H, WRAY N R, GODDARD M E, et al. Estimating missing heritability for disease from genome-wide association studies[J]. American Journal of Human Genetics, 2011, 88(3): 294-305.

[13] HOFFMAN G E. Correcting for population structure and kinship using the linear mixed model: theory and extensions[J]. PloS One, 2013, 8(10): e75707.

[14] IONITA-LAZA I, LEE S, MAKAROV V, et al. Sequence kernel association tests for the combined effect of rare and common variants[J]. The American Journal of Human Genetics, 2013, 92(6): 841-853.

[15] GUSEV A, LEE S H, TRYNKA G, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases[J]. The American Journal of Human Genetics, 2014, 95(5): 535-552.

[16] LU Q, LI B, OU D, et al. A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics[J]. The American Journal of Human Genetics, 2017, 101(6): 939-964.

[17] SIVAKUMARAN S, AGAKOV F, THEODORATOU E, et al. Abundant pleiotropy in human complex diseases and traits[J]. The American Journal of Human Genetics, 2011, 89(5): 607-618.

[18] SIMON K C, MUNGER K L, YANG X, et al. Polymorphisms in vitamin D metabolism related genes and risk of multiple sclerosis[J]. Multiple Sclerosis Journal, 2010, 16(2): 133-138.

[19] MANOUSAKI D, DUDDING T, HAWORTH S, et al. Low-frequency synonymous coding variation in CYP2R1 has large effects on vitamin D levels and risk of multiple sclerosis[J]. The American Journal of Human Genetics, 2017, 101(2): 227-238.

[20] FLETCHER O, HOULSTON R S. Architecture of inherited susceptibility to common cancer[J]. Nature Reviews Cancer, 2010, 10(5): 353.

[21] WANG Q, YANG C, GELERNTER J, et al. Pervasive pleiotropy between psychiatric disorders and immune disorders revealed by integrative analysis of multiple GWAS[J]. Human Genetics, 2015, 134(11-12): 1195-1209.

[22] PANOUTSOPOULOU K, METRUSTRY S, DOHERTY S A, et al. The effect of FTO variation on increased osteoarthritis risk is mediated through body mass index: a Mendelian randomization study[J]. Annals of the Rheumatic Diseases, 2014, 73(12): 2082-2086.

[23] SOLOVIEFF N, COTSAPAS C, LEE P H, et al. Pleiotropy in complex traits: challenges and strategies[J]. Nature Reviews Genetics, 2013, 14(7): 483.

[24] MAURANO M T, HUMBERT R, RYNES E, et al. Systematic localization of common disease-associated variation in regulatory DNA[J]. Science, 2012, 337(6099): 1190-1195.

[25] ALBERT F W, KRUGLYAK L. The role of regulatory variation in complex traits and disease[J]. Nature Reviews Genetics, 2015, 16(4): 197.

[26] FROMER M, P ROUSSOS, SIEBERTS S K, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia[J]. Nature Neuroscience, 2016, 19(11): 1442-1453.

[27] GAGLIANO S A, BARNES M R, WEALE M E, et al. Bayesian method to incorporate hundreds of functional characteristics with association evidence to improve variant prioritization[J]. PLoS One, 2014(9): e98122.

[28] CARBONETTO P, STEPHENS M. Integrated enrichment analysis of variants and pathways in genome-wide association

studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease[J]. PLoS Genetics, 2013, 9(10): e1003770.

[29] MING J, DAI M, CAI M, et al. LSMM: a statistical approach to integrating functional annotations with genome-wide association studies[J]. Bioinformatics, 2018, 34(16): 2788-2796.

[30] ZOU J, HUSS M, ABID A, et al. A primer on deep learning in genomics[J]. Nature Genetics, 2018, 51(1): 12-18.

[31] PARK Y, KELLIS M. Deep learning for regulatory genomics[J]. Nature Biotechnology, 2015, 33(8): 825.

[32] ZHOU J, TROYANSKAYA O G. Predicting effects of noncoding variants with deep learning-based sequence model[J]. Nature Methods, 2015, 12(10): 931.

[33] ERASLAN G, ARLOTH J, MARTINS J, et al. DeepWAS: directly integrating regulatory information into GWASusing deep learning supports master regulator mef2c as risk factor for major depressive disorder[J]. BioRxiv, 2016: 069096.

[34] GAMAZON E R, WHEELER H E, SHAH K P, et al. A gene-based association method for mapping traits using reference transcriptome data[J]. Nature Genetics, 2015, 47(9): 1091.

[35] BARBEIRA A, SHAH K P, TORRES J M, et al. MetaXcan: summary statistics based gene-level association method infers accurate PrediXcan results[J]. BioRxiv, 2016: 045260.

[36] GUSEV A, KO A, SHI H, et al. Integrative approaches for large-scale transcriptome-wide association studies[J]. Nature Genetics, 2016, 48(3): 245.

[37] YANG C, WAN X, LIN X, et al. CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information[J]. Bioinformatics, 2018.

[38] BULIK-SULLIVAN B, FINUCANE H K, ANTTILA V, et al. An atlas of genetic correlations across human diseases and traits[J]. Nature Genetics, 2015, 47(11): 1236.

[39] FURLOTTE N A, ESKIN E. Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model[J]. Genetics, 2015, 200(1): 59-68.

[40] LOH P-R, BHATIA G, GUSEV A, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis[J]. Nature Genetics, 2015, 47(12): 1385-1392.

[41] CHUNG D, YANG C, LI C, et al. GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation[J]. PLoS Genetics, 2014, 10(11): e1004787.

[42] LIU J, WAN X, WANG C, et al. LLR: a latent low-rank approach to colocalizing genetic risk variants in multiple GWAS[J]. Bioinformatics, 2017, 33(24): 3878-3886.

[43] MING J, WANG T, YANG C. LPM: a latent probit model to characterize the relationship among complex traits using summary statistics from multiple GWASs and functional annotations[J]. BioRxiv, 2018: 439133.

[44] KATAN M B. Apolipoprotein E isoforms, serum cholesterol, and cancer[J]. International Journal of Epidemiology, 2004, 33(1): 9.

[45] GRAY R, WHEATLEY K. How to avoid bias when comparing bone marrow transplantation with chemotherapy[J]. Bone Marrow Transplantation, 1991, 7(3): 9-12.

[46] SMITH G D, EBRAHIM S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental

determinants of disease[J]. International Journal of Epidemiology, 2003, 32(1): 1–22.

[47] HEMANI G, BOWDEN J, SMITH G D. Evaluating the potential role of pleiotropy in Mendelian randomization studies[J]. Human Molecular Genetics, 2018, 27(R2): 195–208.

[48] BURGESS S, THOMPSON S G. Interpreting findings from Mendelian randomization using the MR-Egger method[J]. European Journal of Epidemiology, 2017, 32(5): 377–389.

[49] BOWDEN J, SMITH G D, BURGESS S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression[J]. International Journal of Epidemiology, 2015: 44(2): 512–525.

[50] VERBANCK M, CHEN C Y, NEALE B, et al. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases[J]. Nature Genetics, 2018, 50(5), 693.

[51] ZHU Z, ZHENG Z, ZHANG F, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data[J]. Nature Communications, 2018, 9(1): 224.

[52] ZHAO Q, WANG J, HEMANI G, et al. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score[J]. Computer Science, 2018, arXiv:1801.09652.

[53] ZHAO J, MING J, HU X, et al. Bayesian Weighted Mendelian Randomization for Causal Inference based on Summary Statistics[J]. Computer Science, 2018, arXiv:1811.10223.

[54] OKBAY A, BASELMANS B M, DE NEVE J E, et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses[J]. Nature Genetics, 2016, 48(6): 624.

[55] Group of the Psychiatric Genomics Consortium. Identifi cation of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis[J]. The Lancet, 2013, 381(9875): 1371–1379.

[56] XIA J, HE Q, LI Y, et al. The relationship between neuroticism, major depressive disorder and comorbid disorders in Chinese women[J]. Journal of Affective Disorders, 2011, 135(1–3): 100–105.

[57] DE MOOR M H, VAN DEN BERG S M, VERWEIJ K J, et al. Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder[J]. JAMA Psychiatry, 2015, 72(7): 642–650.

[58] KETTUNEN J, DEMIRKAN A, WÜRTZ P, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA[J]. Nature Communications, 2016, 7, 1-9.

[59] LOCKE A E, KAHALI B, BERNDT S I, et al. Genetic studies of body mass index yield new insights for obesity biology[J]. Nature, 2015, 518(7538): 197.

[60] LU Y, DAY F R, GUSTAFSSON S, et al. New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk[J]. Nature Communications, 2016, 7: 10495.

[61] SHUNGIN D, WINKLER T W, CROTEAU-CHONKA D C, et al. New genetic loci link adipose and insulin biology to body fat distribution[J]. Nature, 2015, 518(7538): 187.

[62] VAN DER VALK R J, KREINER-MØLLER E, KOOIJMAN M N, et al. A novel common variant in DCST2 is associated with length in early life and height in adulthood[J]. Human Molecular

Genetics, 2014, 24(4): 1155-1168.
[63] HORIKOSHI M, BEAUMONT R N, DAY F R, et al. Genome-wide associations for birth weight and correlations with adult disease[J]. Nature, 2016, 538(7624): 248-252.
[64] BRADFIELD J P, TAAL H R, TIMPSON N J, et al. A genome-wide association meta-analysis identifies new childhood obesity loci[J]. Nature Genetics, 2012, 44(5): 526.
[65] TAAL H R, ST POURCAIN B, THIERING E, et al. Common variants at 12q15 and 12q24 are associated with infant head circumference[J]. Nature Genetics, 2012, 44(5): 532.
[66] WOOD A R, ESKO T, YANG J, et al. Defining the role of common variation in the genomic and biological architecture of adult human height[J]. Nature Genetics, 2014, 46(11): 1173-1186.
[67] LIU C, KRAJA A T, SMITH J A, et al. Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci[J]. Nature Genetics, 2016, 48(10): 1162-1170.
[68] NIKPAY M, GOEL A, WON H H, et al. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease[J]. Nature Genetics, 2015, 47(10): 1121-1130.
[69] DEN HOED M, EIJGELSHEIM M, ESKO T, et al. Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders[J]. Nature Genetics, 2013, 45(6): 621-631.
[70] NOLTE I M, MUNOZ M L, TRAGANTE V, et al. Genetic loci associated with heart rate variability and their effects on cardiac disease risk[J]. Nature Communications, 2017, 8: 15805.
[71] PATERNOSTER L, STANDL M, WAAGE J, et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis[J]. Nature Genetics, 2015, 47(12): 1449-1456.
[72] LIU J Z, VAN SOMMEREN S, HUANG H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations[J]. Nature Genetics, 2015, 47(9): 979-986.
[73] DUBOIS P C A, TRYNKA G, FRANKE L, et al. Multiple common variants for celiac disease influencing immune gene expression[J]. Nature Genetics, 2010, 42(4): 295-302.
[74] SAWCER S, HELLENTHAL G, PIRINEN M, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis[J]. Nature, 2011, 476(7359): 214-219.
[75] CORDELL H J, HAN Y, MELLS G F, et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways[J]. Nature Communications, 2015, 6(1): 8019.
[76] OKBAY A, BASELMANS B M L, DE NEVE J E, et al. Genetic variants associated with subjective well-being, depressive symptoms and neuroticism identified through genome-wide analyses[J]. Nature Genetics, 2016, 48(6): 624-633.
[77] CENSIN J C, NOWAK C, COOPER N, et al. Childhood adiposity and risk of type 1 diabetes: A Mendelian randomization study[J]. PLoS Medicine, 2017, 14(8): e1002362.
[78] BENTHAM J, MORRIS D L, CUNNINGHAME GRAHAM D S, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus[J]. Nature Genetics, 2015, 47(12): 1457-1464.
[79] MORRIS A P, VOIGHT B F, TESLOVICH T M, et al. Large-scale association analysis provides insights into the genetic

architecture and pathophysiology of type 2 diabetes[J]. Nature Genetics, 2012, 44(9): 981-990.

[80] LU Y, DAY F R, GUSTAFSSON S, et al. New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk[J]. Nature Communications, 2016, 7: 10495.

[81] LI M, LI Y, WEEKS O, et al. SOS2 and ACP1 Loci Identified through Large-Scale Exome Chip Analysis Regulate Kidney Development and Function[J]. Journal of the American Society of Nephrology, 2017, 28(3): 981-994.

[82] STRAWBRIDGE R J, DUPUIS J, PROKOPENKO I, et al. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes[J]. Diabetes, 2011, 60(10): 2624-2634.

[83] TEUMER A, TIN A, SORICE R, et al. Genome-wide association studies identify genetic loci associated with albuminuria in diabetes[J]. Diabetes, 2016, 65(3): 803-817.

[84] PERRY J R, STOLK L, FRANCESCHINI N, et al. Meta-analysis of genome-wide association data identifies two loci influencing age at menarche[J]. Nature Genetics, 2009, 41(6): 648.

[85] MANNING A K, HIVERT M F, SCOTT R A, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance[J]. Nature Genetics, 2012, 44(6): 659-669.

[86] LAMBERT J C, IBRAHIM VERBAAS C A, HAROLD D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease[J]. Nature Genetics, 2013, 45(12): 1452-1458.

[87] BENYAMIN B, HE J, ZHAO Q, et al. Cross-ethnic meta-analysis identifies association of the GPX3-TNIP1 locus with amyotrophic lateral sclerosis[J]. Nature Communications, 2017, 8(1): 611.

[88] PANKRATZ N, BEECHAM G W, DESTEFANO A L, et al. Meta-analysis of Parkinson's disease: identification of a novel locus, RIT2[J]. Annals of Neurology, 2012, 71(3): 370-384.

[89] OTOWA T, HEK K, LEE M, et al. Meta-analysis of genome-wide association studies of anxiety disorders[J]. Molecular Psychiatry, 2016, 21(10): 1391-1399.

[90] SKLAR P, RIPKE S, SCOTT L J, et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4[J]. Nature Genetics, 2011, 43(10): 977-983.

[91] DUNCAN L E, RATANATHARATHORN A, AIELLO A E, et al. Largest GWAS of PTSD (N=20 070) yields genetic overlap with schizophrenia and sex differences in heritability[J]. Molecular Psychiatry, 2017, 23(3): 666.

[92] DUNCAN L, YILMAZ Z, GASPAR H, et al. Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa[J]. American Journal of Psychiatry, 2017, 174(9): 850-858.

[93] PAPPA I, ST POURCAIN B, BENKE K, et al. A genome-wide approach to children's aggressive behavior: the EAGLE consortium[J]. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2016, 171(5): 562-572.

[94] GAO J, DAVIS L K, HART A B, et al. Genome-wide association study of loneliness demonstrates a role for common variation[J]. Neuropsychopharmacology, 2016, 42(4): 811-821.

[95] STEWART S E, YU D, SCHARF J M, et al. Genome-wide association study of obsessive-compulsive disorder[J]. Molecular Psychiatry, 2013, 18(7): 788-798.

作者简介

胡湘红（1991- ），女，深圳市大数据研究院博士生，主要研究方向为生物信息。

彭衡（1974- ），男，香港浸会大学数学系副教授，主要研究方向为金融计量经济学、生物信息、模型选择、非参数方法。

杨灿（1980- ），男，香港科技大学数学系助理教授，主要研究方向为生物信息学、高维数据分析、统计遗传学。

张纵辉（1981- ），男，深圳市大数据研究院副教授，主要研究方向为信号处理、最优化方法、数据通信。

万翔（1972- ），男，深圳市大数据研究院研究科学家，主要研究方向为机器学习、医疗大数据、生物信息。

罗智泉（1963- ），男，深圳市大数据研究院教授，主要研究方向为最优化方法、算法设计、信息科学。

收稿日期：2019-05-07