文章编号:1000-1638(2026)01-0001-12 DOI:10.13484/j.nmgdxxbzk.20260101

应用孟德尔随机化方法识别2型糖尿病的因果基因*

方 舟,许 帅,高 洁,刘俊杰,张利绒
(内蒙古大学物理科学与技术学院,呼和浩特 010021)

摘要:全基因组关联研究(GWAS)能够系统地鉴定与性状和疾病相关的遗传位点。然而,基于GWAS数据识别与疾病显著相关的遗传变异位点和风险基因仍然具有挑战性。本研究通过将GWAS与剪接数量性状位点(sQTL)和表达数量性状位点(eQTL)数据整合,旨在识别2型糖尿病(Type 2 diabetes mellitus,T2DM)的因果基因。基于IEU Open GWAS Project数据库中T2DM的GWAS数据,以及GTEx数据库中49种组织的sQTL和eQTL数据,应用基于汇总数据的孟德尔随机化(SMR)和异质性检验(HEIDI)方法进行共定位分析,鉴定了潜在风险基因,并通过基因的差异表达分析和功能注释探索了其生物学功能。结果表明,4个基因与T2DM的遗传变异显著相关,且在不同组织中发挥重要作用,为T2DM的遗传机制提供了新见解。

关键词:2型糖尿病;因果基因;孟德尔随机化;分子数量性状位点
中图分类号:Q61 **文献标志码**:A

全基因组关联研究(Genome-wide association studies,GWAS)成功地将大量遗传变异与常见疾病以及性状的易感性或特征联系起来,以便更好地了解复杂疾病与性状的遗传机制[1],为提高疾病和性状的生理学见解提供了依据。实际上,确定与疾病相关的生物学路径一直是全基因组关联研究的主要动机[2]。自2002年以来,已经有4 000多个GWAS数据被发表,发现近150 000个标记变异与数百个性状之间存在关联[3]。然而,仅通过GWAS的信号并不能解释每种疾病、每个性状潜在的生物学机制[1],极大地阻碍了我们对复杂性状疾病的遗传认识。那么,是否可以通过可靠的统计方法,联合GWAS寻找疾病的致病基因从而了解致病机制呢?迄今为止,确定非编码遗传变异所调控的靶基因依然是一个亟待解决的关键科学问题。越来越多的研究表明,非编码遗传变异可能通过调节特定基因的mRNA表达水平,在复杂疾病的发生中起到重要作用[4]。利用人类组织的表达数量性状位点(eQTL)数据开展的分析已成为探索疾病潜在机制的重要手段,并在该领域取得了显著进展。有研究证实,基因的可变剪接可能也是一种致病机制[5]。

依据2018年美国糖尿病协会(ADA)颁布的糖尿病分类和诊断标准,糖尿病可分为4种类型:1型糖尿病(Type 1 diabetes mellitus,T1DM)、2型糖尿病(Type 2 diabetes mellitus,T2DM)、妊娠糖尿病和由于其他原因引起的特定类型的糖尿病。T2DM是一种常见的慢性病,对人类的身体健康及劳动力市场造成一定的威胁。2019年全球T2DM患者为4.63亿,至2045年预计将达到7亿,约增长51%。国际糖尿病联盟(IDF)流行病学调查显示,占全球糖尿病病例超过90%的T2DM,其发病机制呈现多因素

* **收稿日期**:2025-07-17;**修回日期**:2025-11-04
基金项目:国家自然科学基金项目(61962041);内蒙古自治区自然科学基金项目(2024MS03023)
作者简介:方 舟(2000—),女,内蒙古乌兰察布人,2022级硕士研究生。E-mail:1874056406@qq.com
通信作者:张利绒(1972—),女,内蒙古乌兰察布人,教授,博士。主要从事理论生物物理研究。E-mail:pyzlr@imu.edu.cn

协同作用特征。根据ADA数据,在2017—2018年,美国青少年糖尿病的发病数约为18 200例T1DM,5 300例T2DM。我国糖尿病类型以T2DM为主,IDF调查显示糖尿病人群中T2DM占90%以上,T1DM和其他类型糖尿病少见。2015—2019年,我国T2DM的总体患病率已达到14.92%。

为深入挖掘T2DM的致病机制,基于IEU Open GWAS Project数据库中T2DM的GWAS数据,联合GTEx数据库中49种组织的剪接数量性状位点(sQTL)和表达数量性状位点(eQTL)数据,应用基于汇总数据的孟德尔随机化(Summary-data-based Mendelian randomization,SMR)和异质性检验(Heterogeneity in dependent instruments,HEIDI)方法,识别了因sQTL和eQTL多效性而与T2DM显著相关的基因和遗传变异位点。

1 数据和方法

1.1 数据来源

本文以T2DM为研究对象,从IEU Open GWAS Project数据库(https://gwas.mrcieu.ac.uk/datasets/ebi-a-GCST006867/)下载了2018年发表的T2DM的GWAS汇总统计数据。该数据基于欧洲人群的61 714例T2DM病例和1 178例对照组,对5 030 727个SNP(HG19/GRCh37)进行关联分析。在GEO数据库(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE280402)下载了T2DM的RNA-seq数据(GSE280402),包括8个患病样本和8个正常样本。从eQTLGen数据库(https://eqtlgen.org/cis-eqtls.html)下载了血液的cis-eQTL汇总统计数据,样本量为31 684,SNP个数为10 525。从GTEx数据库(https://yanglab.westlake.edu.cn/data/SMR/GTEx_V8_cis_sqtl_summary.html)下载了V8发布的49个人体组织的cis-sQTL与cis-eQTL汇总统计数据[6],样本量范围为73~670例,sQTL和eQTL位于基因转录起始位点(TSS)上游和下游1 Mb内的SNP,且显著性水平$P<5\times10^{-8}$。

1.2 方法

孟德尔随机化(Mendelian randomization,MR)是一种受自然随机化实验启发而产生的因果推断方法,其核心在于利用遗传变异(通常是单核苷酸多态性,SNP)作为工具变量(Instrumental variable,IV),来评估暴露因素(如基因表达)与结局(如疾病)之间的潜在因果关系。该方法能够有效规避传统观察性研究中的混杂偏倚和反向因果关系,其有效性建立在3个核心假设之上:(1)关联性假设,工具变量必须与暴露因素(如基因表达水平)强相关;(2)独立性假设,工具变量必须与影响暴露和结局的混杂因素无关;(3)排他性假设,工具变量只能通过影响暴露因素来间接影响结局,而不能存在其他直接或间接的路径,如图1所示。

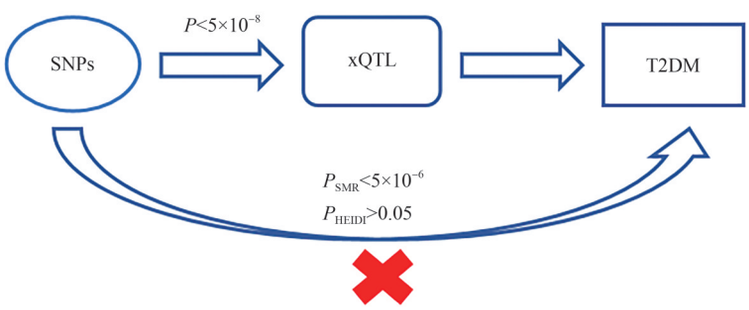


图1 孟德尔随机化研究核心假设

Fig. 1 Core assumptions of Mendelian randomization study

文献[7-8]提出SMR&HEIDI方法(http://cnsgenomics.com/software/smr/),使用来自GWAS和分子数量性状位点(xQTL)研究的汇总统计数据,旨在鉴定基因表达水平、可变剪接等分子表型与复杂性状之间的多效性关联。SMR&HEIDI方法由两部分构成,分别是基于汇总数据的孟德尔随机化方法(SMR)和异质性检验(HEIDI)分析。SMR分析基于GWAS汇总数据,旨在发现因果变异对某一性状

的影响是否是通过基因表达水平以及可变剪接等方式介导的，其分析原理与MR类似，将MR的测试扩展到xQTL与GWAS的因果关系鉴定中，能够定量给出xQTL对疾病与性状的效应值。HEIDI检验通过检测基因型与表型关联的异质性来区分多效性，当$P_{HEIDI}>0.05$时，表明观察到的关联更可能由单一因果变异驱动。

为满足核心假设，设定SNP纳入标准：(1)cis-xQTL优先原则，严格筛选位于基因转录起始位点上下游1 Mb范围内的cis-xQTLs；(2)显著性水平$P<5\times10^{-8}$；(3)次要等位基因频率>0.01。同时采用HEIDI检验区分因果关联并排除多效性，当$P_{HEIDI}\leqslant0.05$时表明观测到的关联可能源于高度连锁不平衡的两个独立遗传变异。最终，符合上述标准的SNP将作为工具变量，确保基因表达或可变剪切与T2DM风险之间的因果关系不受混杂因素影响。

2　结果

2.1　T2DM中与组织特异性表达相关的因果基因

2.1.1　整合GWAS与组织sQTL推断因果基因

首先利用SMR方法在GTEx数据库的49种组织中广泛筛选与T2DM存在潜在因果关系的基因和SNP，旨在获得一个全面的T2DM因果基因初始集合，并为后续分析奠定基础。

本文对GTEx数据库下载的49种组织的cis-sQTL与T2DM的GWAS汇总数据进行了SMR分析。在46种组织中，共得到415个gene-sQTL对($P_{SMR}<5\times10^{-6}$，$P_{HEIDI}>0.05$)，对应了49个潜在风险基因和110个SNP位点。其中有9个基因为人类白细胞抗原(Human leukocyte antigen，HLA)类基因，在免疫系统中起着关键作用，主要参与抗原呈递和免疫反应的调节。近年来，研究发现HLA类基因与T2DM之间也存在一定的关联，尤其是在自身免疫反应、疾病进展和并发症方面[9]。在3种组织，即脑尾状基底节(Brain caudate basal ganglia)组织、大脑皮质组织(Brain cortex)和脑壳核基底节(Brain putamen basal ganglia)中未发现因果基因，结果见图2。

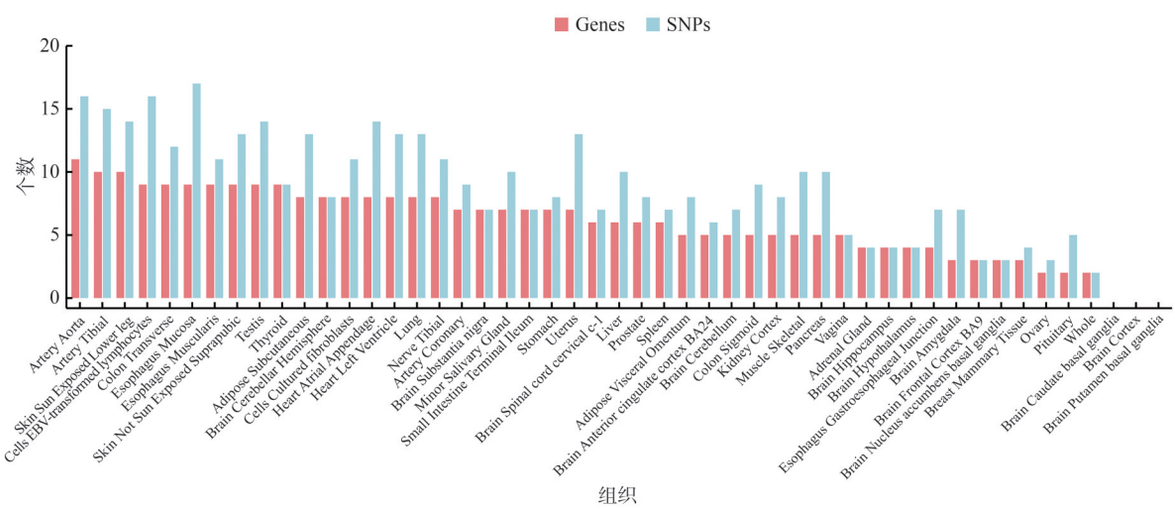


图2　49种组织中T2DM相关基因和SNP的频次

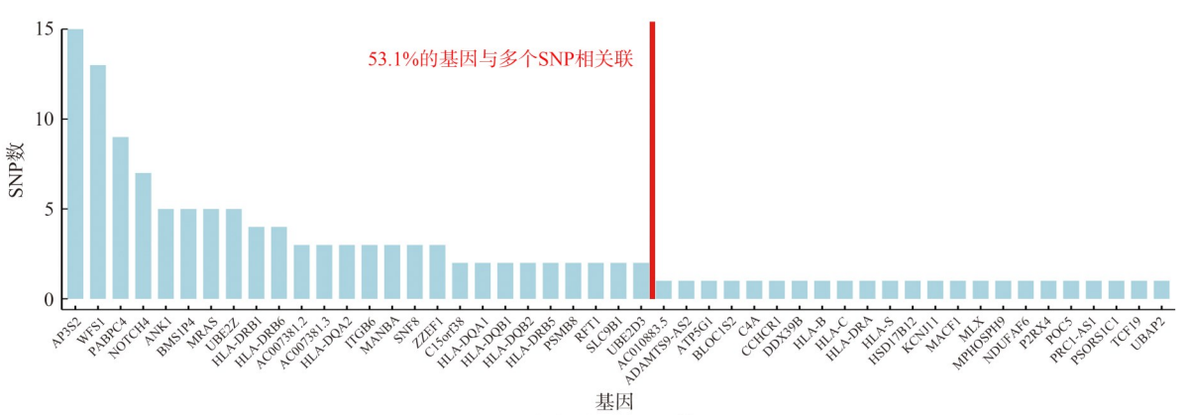
Fig. 2　Frequencies of genes and SNPs associated with T2DM in 49 tissues

在这些因果基因中，存在一个基因对应多个SNP的现象，图3(a)给出了基因对应SNP数量的分布。我们发现，53.1%的基因与多个SNP相关联，如基因*AP3S2*对应15个SNP，说明该基因的剪接可能通过不同遗传变异共同调控。通过统计剪接风险基因在组织中出现的频次，如图3(b)所示，发现61.3%的基因在多个组织出现，如基因*AP3S2*在27种组织中出现，该基因广泛调控着不同组织中基因的剪接。文献[9]发现多项GWAS研究均证明了在不同人群中发现了基因*AP3S2*在多种组织中与T2DM的关联，且与rs4932265位点相关，验证了我们的结论。

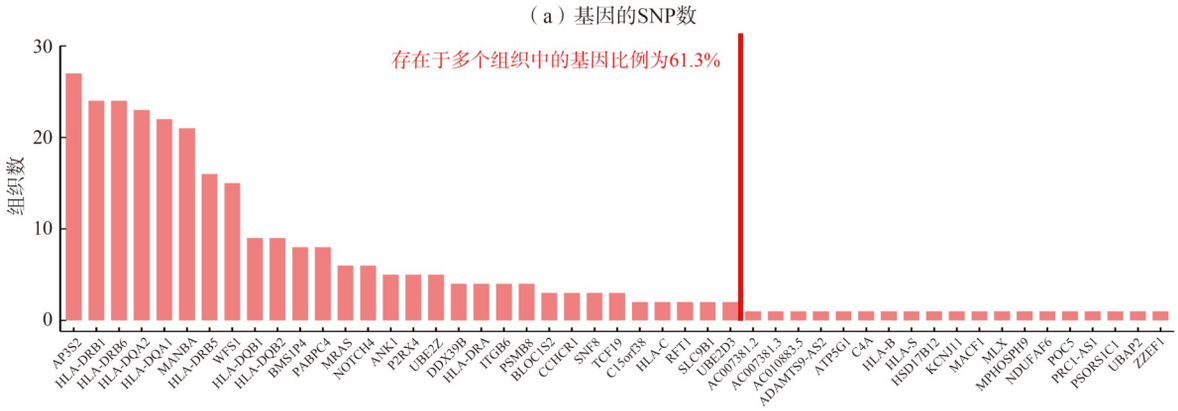
在415个gene-sQTL对中,对应了110个SNP。统计发现,存在一个SNP对应多个基因的现象,图3(c)给出了SNP对应基因的数量分布。在110个SNP中,11.9%的SNP与多个基因相关联,如位点rs9271775对应了5个基因,表明该位点可能同时调控多个基因的剪接。最后,通过统计剪接风险SNP在组织中出现的频次,如图3(d)所示,发现32.8%的SNP在多个组织中出现,如位点rs660895在24种组织中出现,意味着该遗传变异广泛调控着不同组织中基因的可变剪切。

2.1.2 整合GWAS与组织eQTL推断因果基因

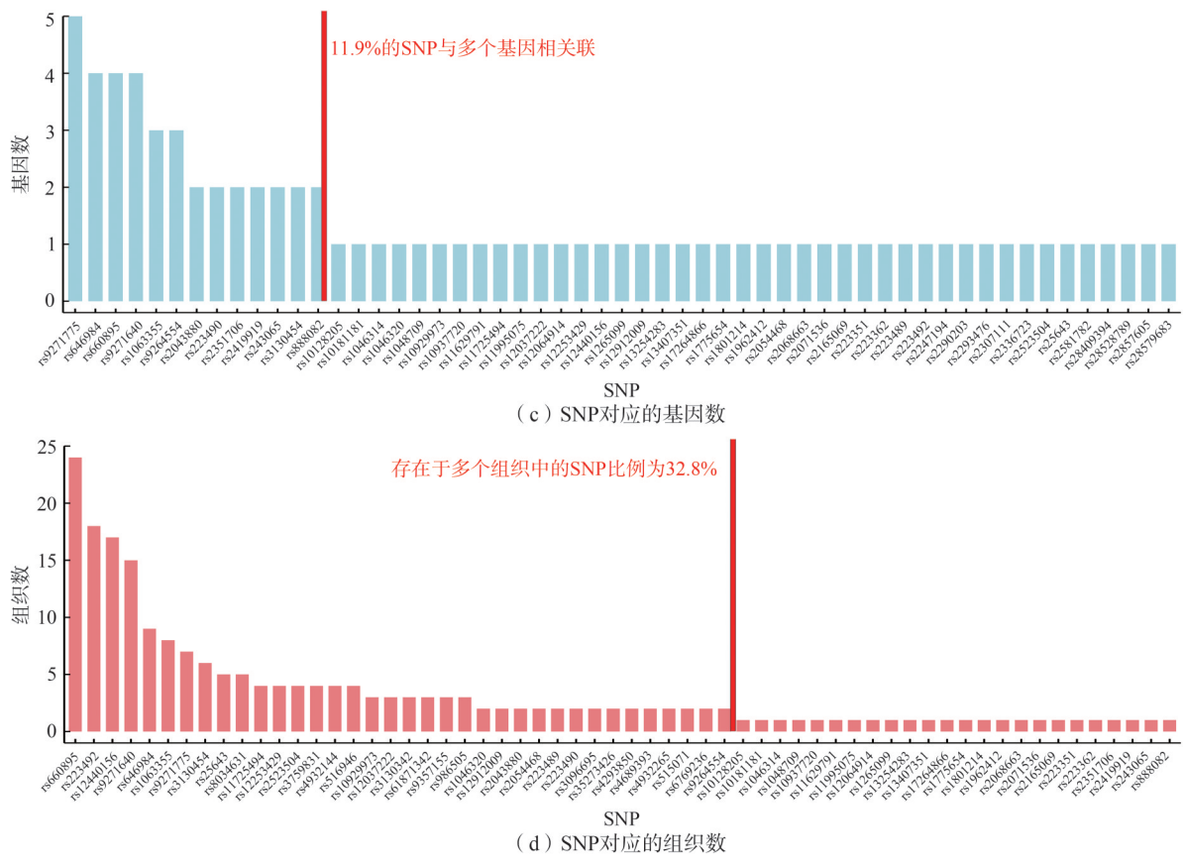
SNP对T2DM产生影响的另一个可能途径是通过介导靶基因的表达来影响性状。将GWAS与GTEx数据库中49种组织的eQTL数据整合,进行了SMR分析。同样,仅关注了GTEx数据库中的cis-eQTL数据,即位于基因转录起始位点(TSS)上游和下游1 Mb距离,且与T2DM显著相关的SNP($P<5\times10^{-8}$)。应用SMR&HEIDI方法,检测到437个gene-eQTL对($P_{SMR}<5\times10^{-6}$,$P_{HEIDI}>0.05$),对应了84个潜在风险基因和156个SNP位点。其中,有些基因对应了多个SNP。图4(a)显示了基因与SNP数量的分布情况,在84个基因中,有44.7%的基因与多个SNP相关联,如基因*PABPC4*对应9个SNP,说明该基因的表达可能受到不同遗传变异的共同调控。通过统计表达特异风险基因在组织中出现的频次,如图4(b)所示,发现59.6%的基因在多个组织中频繁出现,如基因*CWF19L1*在38种组织中出现,意味着该gene-SNP对中的遗传变异广泛调控着不同组织中基因的表达。在437个gene-eQTL对中,对应了156个SNP。统计发现,存在一个SNP对应多个基因的现象,图4(c)显示了SNP与基因数量的分布情况,在156个SNP中,有12.2%的SNP与多个基因相关联,如位点rs1063355对应了9个基因,表明该位点可能同时调控多个基因的表达。最后,通过统计表达特异性风险SNP出现的频次,如图4(d)所示,发现35.9%的SNP在多个组织中出现,如位点rs1063355在36种组织中出现,意味着该遗传变异广泛调控着不同组织中基因的表达。



(a)基因的SNP数



(b)基因对应的组织数



（c）SNP对应的基因数

（d）SNP对应的组织数

图 3　剪接相关基因和 SNP 位点的频次分布

Fig. 3　Frequency distributions of splicing-associated genes and SNPs

有研究表明，基因 *PABPC4* 作为中介因子，受特定 DNA 甲基化位点（cg15123755）的调控，其表达量的增加能通过提升高密度脂蛋白胆固醇水平来显著降低患 2 型糖尿病的风险[10]。T2DM 患者常伴随慢性低度炎症和胰岛素抵抗，而 *CWF19L1* 可以通过调节免疫相关基因的表达，来影响炎症反应和代谢稳态[11]。研究证明，rs1063355 通过调节基因 *HLA-DQB1* 的表达水平，影响免疫反应和代谢稳态[12-13]，从而与 T2DM 的发病机制相关，进一步证明了结果的可靠性。

在 GWAS 与 eQTL 分析中，皮下脂肪与甲状腺组织对应的基因和 SNPs 均最高，表明其基因表达受遗传变异影响显著；肾脏皮质组织无对应的基因与 SNP，可能因研究样本不足所致，如图 5 所示。总体来看，多数组织的 eQTL 基因和 SNP 数较高，表明基因表达受遗传变异广泛影响，体现了 eQTL 调控的普遍性。eQTL 活跃性在脂肪和甲状腺等组织中更高，可能与组织复杂性或研究深度有关，体现了 eQTL 调控的组织特异性。

图 2—4 展示了通过 SMR 方法初步筛选出的与 T2DM 潜在相关的 xQTLs 在不同组织中的分布概况，其结果揭示了 T2DM 遗传基础的组织特异性和复杂性，为我们后续聚焦于在多种组织中重复出现、可能具有核心功能的基因提供了初步线索和筛选依据。

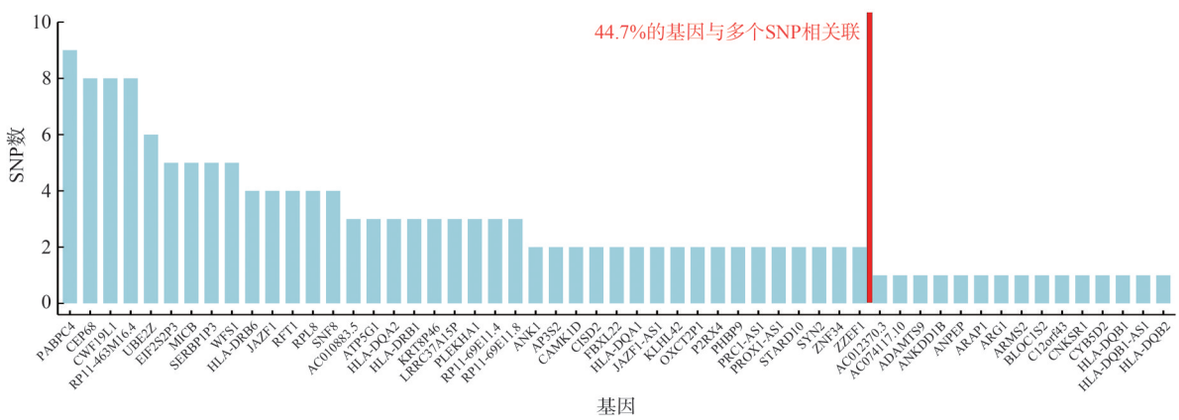
2.1.3　整合 GWAS 与血液 eQTL 推断因果基因

在 eQTLGen 数据库中，下载血液的 cis-eQTL 数据，将满足 $P<5\times10^{-8}$ 的 eQTL 作为研究对象，应用 SMR&HEIDI 方法检测到 24 个 gene-eQTL 对（$P_{SMR}<5\times10^{-6}$，$P_{HEIDI}>0.05$），对应了 24 个 T2DM 潜在因果基因与 24 个 SNP。其中，5 个基因与 49 种组织 eQTL 数据的 Whole 组织鉴定得到的基因重复，分别是 *MAP3K13*、*RPL8*、*ARG1*、*CWF19L1*、*PABPC4*，结果如图 6 所示。文献[14]调研发现 *MAP3K13* 的表达在多种组织中广泛存在，包括肾上腺和肾脏，这可能与 T2DM 的代谢紊乱有关。*RPL8* 的表

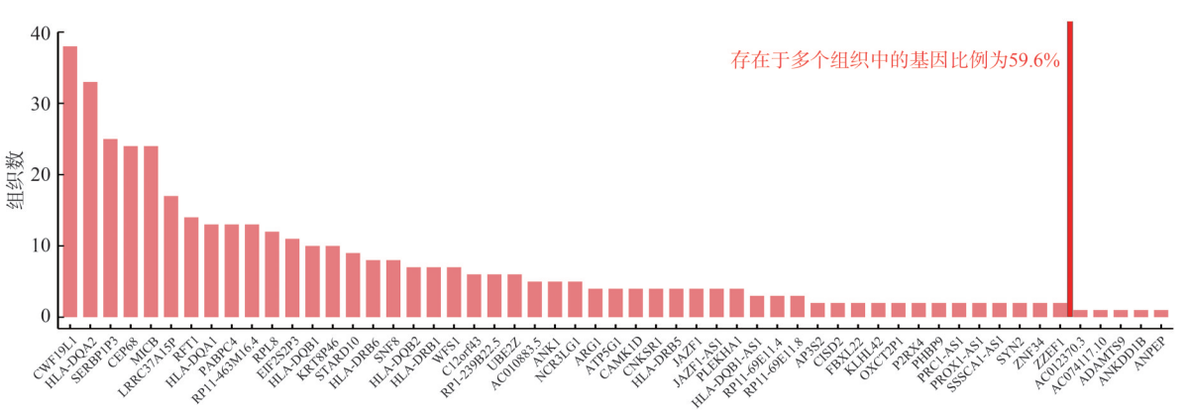
达变化可能通过影响胰岛β细胞的蛋白质合成和功能，间接影响T2DM的发病机制[15]。*ARG1*可能通过调节氮代谢和炎症反应影响T2DM的发病机制，此外，在脂肪组织和肝脏中的表达可能与脂质代谢和胰岛素敏感性有关，进一步影响T2DM的进展[16]。

2.2 组织特异性差异表达因果基因的注释

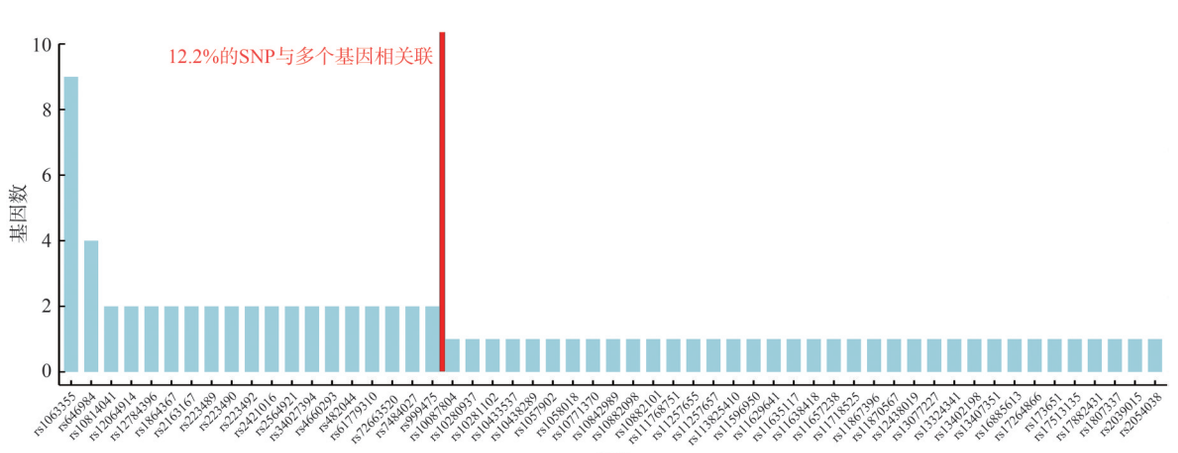
为进一步鉴别出最可能受疾病状态影响的核心基因，基于上述跨组织筛选出的T2DM因果基因初始集合，利用来自GEO数据库的RNA-seq转录组数据，通过差异表达分析对初始基因集合进行验证和精细化筛选。



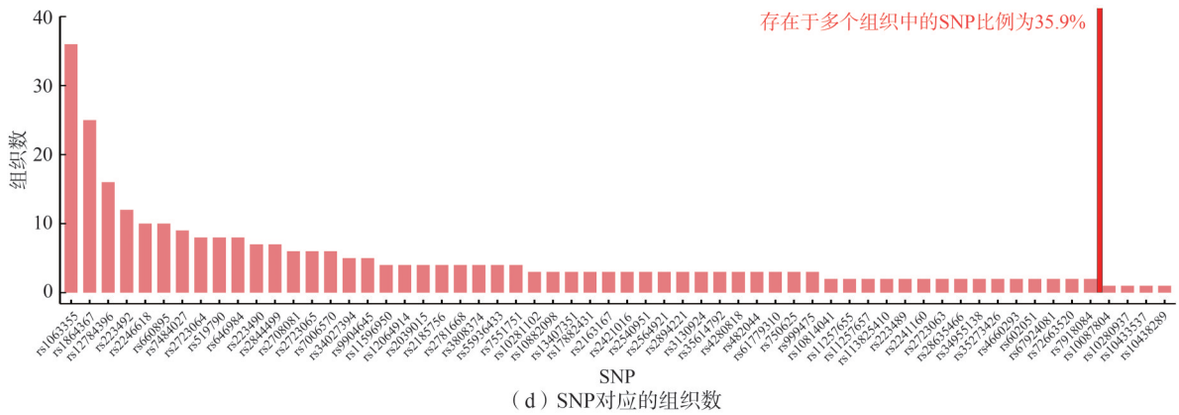
（a）基因的SNP数



（b）基因对应的组织数



（c）SNP对应的基因数



（d）SNP对应的组织数

图4　表达相关基因和SNP的频次分布

Fig. 4　Frequency distributions of expression-associated genes and SNPs

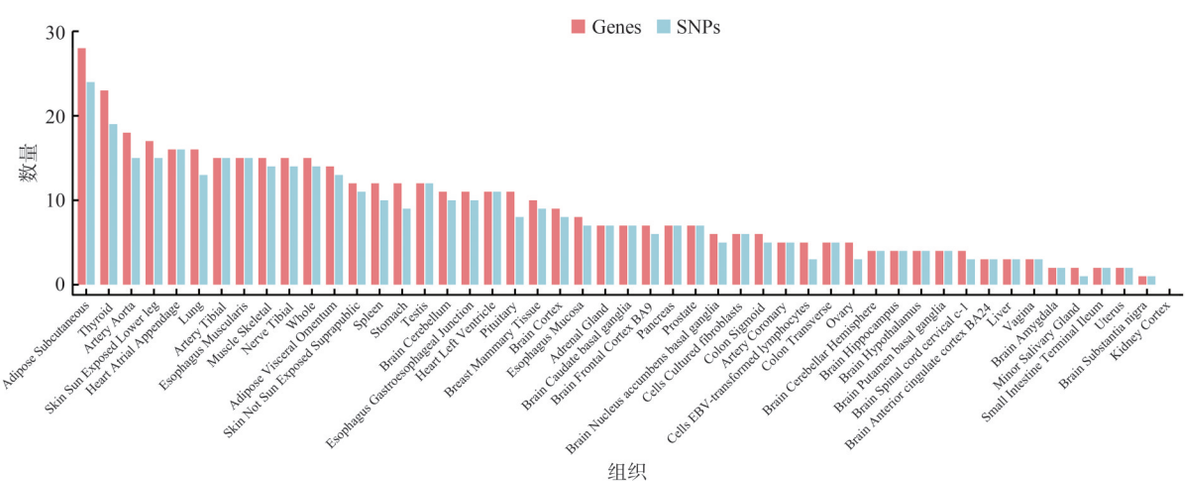


图5　49种组织对应的表达基因和SNP

Fig. 5　Expressed genes and SNPs associated with 49 tissues

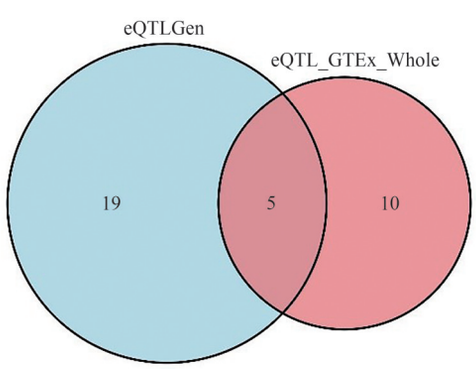


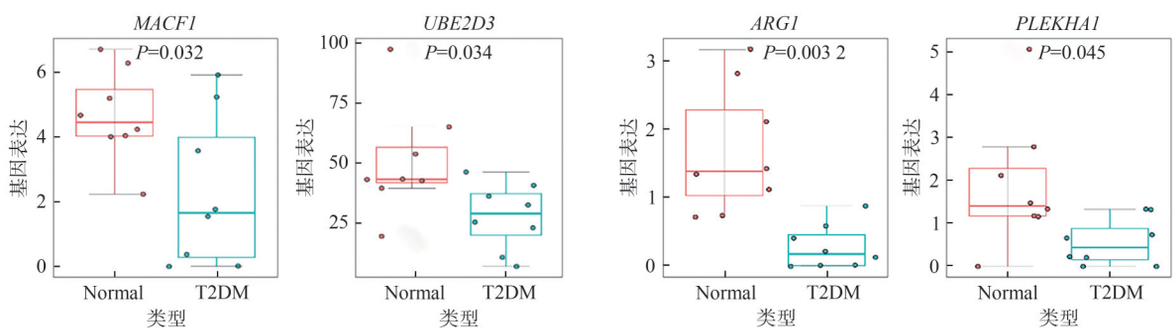
图6　两组血液eQTL联合分析的靶基因

Fig. 6　Overlap of target genes derived from association analysis of two blood eQTL

在49种组织的GWAS-sQTL分析中得到的415个gene-sQTL对中，包括49个基因，110个SNP，其中43个基因有匹配的RNA-seq数据（GSE280402）。针对这43个基因，使用独立样本t检验的方法，比较了其在T2DM和正常样本中的表达水平。在进行t检验前，使用Shapiro-Wilk检验和Levene检验分别验证了基因表达数据的正态性（$P>0.05$）与方差齐性（$P>0.05$）。经检验，共有13个基因的数据同时满足这两项参数检验的前提假设，样本均为独立采集，满足独立性假设。在此基础上进行的

独立样本 t 检验结果显示,*MACF1*(P=0.032)、*UBE2D3*(P=0.034)基因的表达在 T2DM 中显著下调,见图 7(a)。这一发现提示 *MACF1* 和 *UBE2D3* 可能参与了 T2DM 的发生发展过程。其中,*MACF1* 中的 *M2290V* 变体被证明会增加患 T2DM 的风险[17]。而 *UBE2D3* 通过泛素化调控 p53 等关键分子,参与细胞凋亡抑制、DNA 损伤修复及癌症进展,其表达下调可能会促进 β 细胞凋亡与胰腺重塑[18]。

在 49 种组织的 GWAS-eQTL 分析中,得到了 437 个 gene-eQTL 对,对应于 84 个基因,156 个 SNP。其中 69 个基因有 RNA-seq 数据(GSE280402)。针对这 69 个基因,使用独立样本 t 检验的方法,比较了其在 T2DM 和正常样本中的表达水平。在进行 t 检验前,使用 Shapiro-Wilk 检验和 Levene 检验分别验证了基因表达数据的正态性(P>0.05)与方差齐性(P>0.05)。经检验,共有 25 个基因的数据同时满足这两项参数检验的前提假设,样本均为独立采集,满足独立性假设。在此基础上进行的独立样本 t 检验结果显示,*ARG1*(P=0.003 2)、*PLEKHA1*(P=0.045)基因的表达在 T2DM 中显著下调,可能与 T2DM 的发病机制有关,结果如图 7(b)所示。已有研究表明,在 T2DM 病理状态下,红细胞衍生的细胞外囊泡可通过转运 ARG1 至内皮细胞,诱导氧化应激反应,从而导致内皮功能障碍[19]。另一方面,PLEKHA1 作为重要的信号调控分子,能够通过其 C 端 PH 结构域特异性结合 PtdIns(3,4)P_2,进而调节 PtdIns(3,4,5)P_3水平,最终通过激活 Akt 信号通路影响胰岛素敏感性[20]。



(a) sQTL显著差异表达的基因 (b) eQTL显著差异表达的基因

图 7 差异表达基因的表达水平对比

Fig. 7 Expression level comparison of differentially expressed genes

通过独立样本 t 检验的方法得到了 4 个关键基因,利用 GeneCards 数据库查找这 4 个关键基因的生物学功能,其具体相关功能如表 1 所示。

2.3 因果基因功能分析

2.3.1 GO 和 KEGG 分析

GO 富集分析阐明了目标基因集在特定生物学功能上的富集情况[21],而 KEGG 富集分析可以说明基因和功能通路之间的关系[22]。本文运用 GO 和 KEGG 方法对 49 种组织与 T2DM 的差异基因集合进行关键基因富集分析,旨在了解关键基因的相关功能以及代谢途径等信息[23]。

如图 8 和 9 所示,综合 GO 与 KEGG 结果可知,基因 *ARG1*、*MACF1* 和 *UBE2D3* 在多种生物学过程和代谢通路中发挥了重要作用,这些过程和通路与 T2DM 的发病机制存在潜在联系。在代谢调节方面,基因 *ARG1* 参与了氨基酸生物合成和精氨酸与脯氨酸代谢,这些过程对维持氮平衡和整体代谢状态至关重要,代谢紊乱可能导致胰岛素抵抗,进而影响 T2DM 的发展;基因 *MACF1* 则通过调节谷胱甘肽代谢影响细胞内的氧化还原状态,氧化应激的增加可能损害胰岛素信号传导和敏感性[24]。*UBE2D3* 在调控线粒体蛋白定位通路中的富集,揭示了其在维持线粒体代谢核心功能方面可能扮演着重要角色,*UBE2D3* 的功能失调可能会影响 T2DM 的发展。此外,基因 *MACF1* 在造血细胞谱系和肾素-血管紧张素系统中的功能,可能通过影响免疫系统功能和血压调节间接影响 T2DM 的发病机制[25]。总之,基因 *ARG1*、*MACF1* 和 *UBE2D3* 通过代谢重编程(氨基酸、谷胱甘肽)和线粒体功能调控双重机制参与 T2DM 的发生发展。未来可结合单细胞测序和空间转录组,细化这些基因在胰岛

β细胞和脂肪组织等微环境中的时空表达特征，并探索其作为治疗靶点的潜力（如*ARG1*抑制剂或*MACF1*抗氧化调节剂）。

表 1　差异表达相关基因的功能注释

Table 1　Functional annotation of differentially expressed genes

项目	*MACF1*	*UBE2D3*	*ARG1*	*PLEKHA1*
xQTL-T2DM	GWAS-sQTL	GWAS-sQTL	GWAS-eQTL	GWAS-eQTL
SNP数	1	2	1	3
组织数	1	2	4	4
基因表达	下调	下调	下调	下调
功能	*MACF1*通过整合细胞骨架动力学与Wnt信号，在细胞运动、极化和代谢调控中起枢纽作用，其异常可能间接促进T2DM的发生发展	*UBE2D3*（P61077）编码E2泛素结合酶，通过催化K11/K48多泛素化及单泛素化修饰，参与p53调控、DNA修复、EGFR内吞、蛋白质稳态、抗病毒免疫（RIG-I）及核糖体质量控制等过程，调控细胞应激与信号传导	*ARG1*基因编码的Ⅰ型精氨酸酶通过调控尿素循环和免疫微环境中的精氨酸代谢，在氮代谢平衡和免疫抑制中发挥双重作用，可能间接影响2型糖尿病（T2DM）的发生发展	*PLEKHA1*基因编码一种含有pleckstrin同源结构域的衔接蛋白质。编码的蛋白质定位于质膜，并特异性结合磷脂酰肌醇3,4-二磷酸，这种蛋白质可能参与质膜中信号复合物的形成

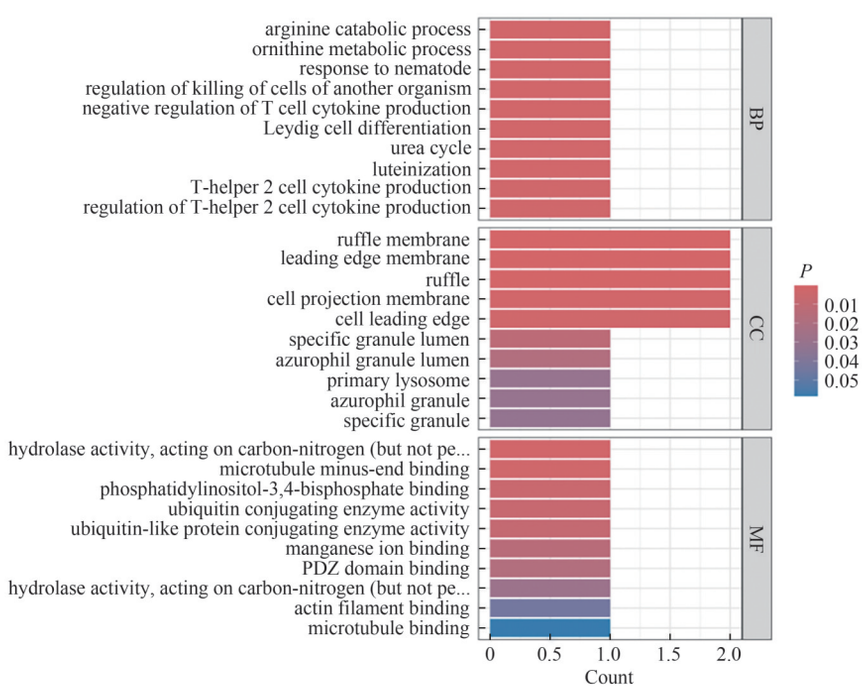


图 8　差异表达基因的GO分析

Fig. 8　GO analysis of differentially expressed genes

2.3.2　因果基因的PPI分析

运用STRING数据库[26]以及GeneMANIA在线工具绘制了蛋白互作（PPI）网络。基于STRING数据库构建了T2DM相关蛋白质相互作用网络，结果如图10（a）所示，涵盖49种组织中的4个关键基因。采用默认参数设置，所得网络经PPI富集分析显示极显著富集（$P<1\times10^{-16}$），包含34个节点（含4个关键基因）和98条相互作用边，揭示了核心基因与其他蛋白质的紧密关联。

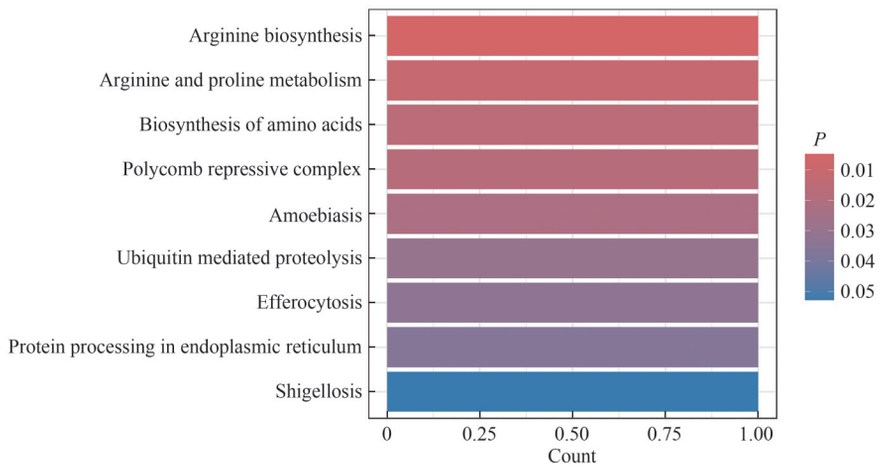
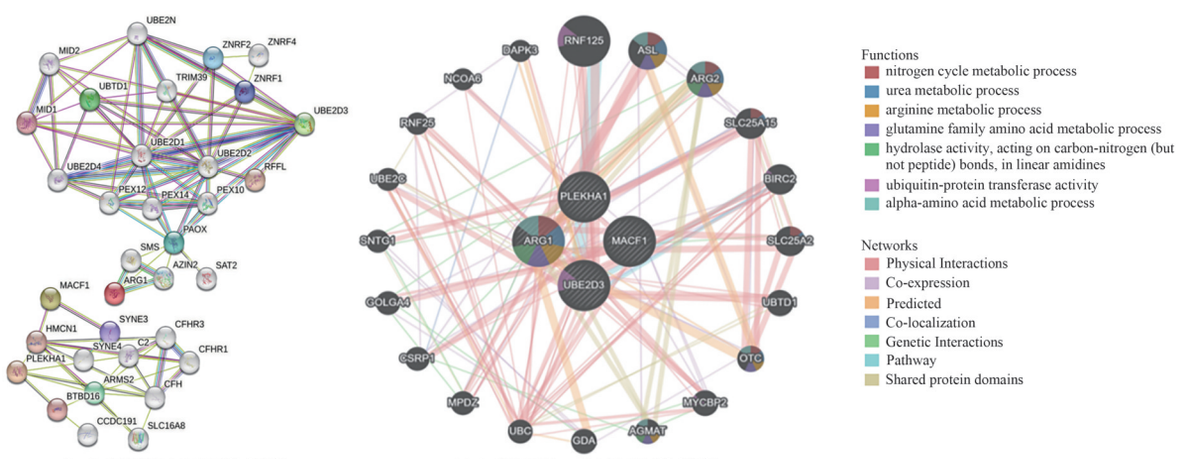


图 9 差异表达基因的 KEGG 分析

Fig. 9 KEGG analysis of differentially expressed genes



(a) 基于STRING数据库 (b) 基于GeneMANIA数据库

图 10 差异表达基因的 PPI 分析

Fig. 10 PPI analysis of differentially expressed genes

进一步,通过 GeneMANIA 平台拓展网络构建,共纳入 20 个潜在相互作用基因,形成包含 4 个关键基因在内的 24 个节点和 158 条相互作用边,见图 10(b)。相互作用类型通过颜色编码呈现:红色(77.64%)代表物理相互作用,紫色(8.01%)指示共表达关联,橙色(5.37%)表征其他功能关联,其余部分可能涉及预测或文献支持等机制。节点的颜色梯度映射基因功能,系统阐释了核心基因在 T2DM 病理网络中的调控枢纽作用及其潜在生物学功能。

以上结果与之前的 GO、KEGG 分析相似,进一步验证了之前结果的可信性。

3 结论

本文基于 IEU Open GWAS Project 数据库中 T2DM 的 GWAS 数据和 GTEx 数据库中 49 种组织的 sQTL 和 eQTL 数据,利用 SMR&HEIDI 方法识别了 49 种组织的剪接、基因表达与 T2DM 的关键基因,构建了 49 种组织的 sQTL 和 eQTL 基因集合,统计了不同基因对应的 SNP 数量以及在不同组织中的出现频次。然后,基于 GEO 数据库 T2DM 的 RNA-seq 数据,分析了 T2DM 和正常样本基因的表达水平,得到了 T2DM 的差异表达基因,构建了差异表达基因集合。将 49 种组织的 sQTL 和 eQTL 中基因的表达数据从 T2DM 的 RNA-seq 数据中提取出来,使用独立样本 t 检验的方法,得到了 4 个关键基因,并对由这 4 个基因构成的基因集合进行了基因功能分析、GO 和 KEGG 分析以及 PPI

网络分析。在之后的工作中，可以对已经确认的基因进行分子对接和模拟分析，进一步研究相应的靶向药物，为 T2DM 的治疗提供一定的理论依据。

参考文献：

[1] BROEKEMA R V，BAKKER O B，JONKERS I H. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era[J]. Open Biology，2020，10(1)：190221.

[2] REYNOLDS R H，HARDY J，RYTEN M，et al. Informing disease modelling with brain-relevant functional genomic annotations[J]. Brain，2019，142(12)：3694-3712.

[3] BUNIELLO A，MACARTHUR J A L，CEREZO M，et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies，targeted arrays and summary statistics 2019[J]. Nucleic Acids Research，2019，47(D1)：D1005-D1012.

[4] GREEN J，CAIRNS B J，CASABONNE D，et al. Height and cancer incidence in the million women study：Prospective cohort，and meta-analysis of prospective studies of height and total cancer risk[J]. The Lancet Oncology，2011，12(8)：785-794.

[5] QI T，WU Y，FANG H L，et al. Genetic control of RNA splicing and its distinct role in complex trait variation[J]. Nature Genetics，2022，54(9)：1355-1363.

[6] HAMILTON M T，HAMILTON D G，ZDERIC T W. Role of low energy expenditure and sitting in obesity，metabolic syndrome，type 2 diabetes，and cardiovascular disease[J]. Diabetes，2007，56(11)：2655-2667.

[7] ZHU Z H，ZHANG F T，HU H，et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets[J]. Nature Genetics，2016，48(5)：481-487.

[8] WU Y，ZENG J，ZHANG F T，et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits[J]. Nature Communications，2018，9(1)：918.

[9] BOULAND G A，BEULENS J W J，NAP J，et al. Diabetes risk loci-associated pathways are shared across metabolic tissues[J]. BMC Genomics，2022，23(1)：368.

[10] NIKPAY M. Genome-wide search identified DNA methylation sites that regulate the metabolome[J]. Frontiers in Genetics，2023，14：1093882.

[11] ZHANG Y Q，YI J J，WEI G G，et al. *CWF19L1* promotes T-cell cytotoxicity through the regulation of alternative splicing[J]. Journal of Biological Chemistry，2024，300(12)：107982.

[12] SHU X，PURDUE M P，YE Y Q，et al. Multilevel-analysis identify a cis-expression quantitative trait locus associated with risk of renal cell carcinoma[J]. Oncotarget，2015，6(6)：4097-4109.

[13] 孙肖霄，黄干，谢志国，等. 1型糖尿病遗传学研究进展[J]. 中华医学杂志，2020，100(10)：793-796.

[14] DAFTUAR L，ZHU Y，JACQ X，et al. Ribosomal proteins RPL37，RPS15 and RPS20 regulate the Mdm2-p53-MdmX network[J]. PLoS One，2013，8(7)：e68667.

[15] XU L L，YANG G，SONG B，et al. Ribosomal protein L8 regulates the expression and splicing pattern of genes associated with cancer-related pathways[J]. Turkish Journal of Biology，2023，47(5)：313-324.

[16] BENAVENTE M A，BIANCHI C P，ABA M A. Expression of arginine vasopressin type 2 receptor in canine mammary tumours：Preliminary results[J]. Journal of Comparative Pathology，2020，179：36-40.

[17] ALBRECHTSEN A，GRARUP N，LI Y，et al. Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes[J]. Diabetologia，2013，56(2)：298-310.

[18] LEI Z，JUNHUI L，PEIFENG L. Candidate genes mediated by estrogen-related receptor γ in pancreatic β cells[J]. Journal of Biochemical and Molecular Toxicology，2019，33(10)：e22390.

[19] COLLADO A，HUMOUD R，KONTIDOU E，et al. Erythrocyte-derived extracellular vesicles induce endothelial dysfunction through arginase-1 and oxidative stress in type 2 diabetes[J]. The Journal of Clinical Investigation，2025，135(10)：e180900.

[20] HU Y，TAN L J，CHEN X D，et al. Identification of novel variants associated with osteoporosis，type 2 diabetes and potentially pleiotropic loci using pleiotropic cFDR method[J]. Bone，2018，117：6-14.

[21] KONDO T, NAKANO Y, ADACHI S, et al. Effects of tobacco smoking on cardiovascular disease[J]. Circulation Journal: Official Journal of the Japanese Circulation Society, 2019, 83(10): 1980-1985.
[22] LU Y L, WANG Z, ZHENG L R. Association of smoking with coronary artery disease and myocardial infarction: A Mendelian randomization study[J]. European Journal of Preventive Cardiology, 2021, 28(12): e11-e12.
[23] MANRIQUE-GARCIA E, SIDORCHUK A, HALLQVIST J, et al. Socioeconomic position and incidence of acute myocardial infarction: A meta-analysis[J]. Journal of Epidemiology and Community Health, 2011, 65(4): 301-309.
[24] ROBERTS B A, BATTY G D, GALE C R, et al. IQ in childhood and atherosclerosis in middle-age: 40 Year follow-up of the Newcastle thousand families cohort study[J]. Atherosclerosis, 2013, 231(2): 234-237.
[25] NASS O, HOFF D A, LAWLOR D, et al. Education and adult cause-specific mortality: Examining the impact of family factors shared by 871 367 Norwegian siblings[J]. International Journal of Epidemiology, 2012, 41(6): 1683-1691.
[26] KHAN S S, NING H Y, WILKINS J T, et al. Association of body mass index with lifetime risk of cardiovascular disease and compression of morbidity[J]. JAMA Cardiology, 2018, 3(4): 280-287.

Systematic Identification and Characterization of Causal Risk Genes for T2DM Using Mendelian Randomization Methods

FANG Zhou, XU Shuai, GAO Jie, LIU Junjie, ZHANG Lirong
(*School of Physical Science and Technology*, *Inner Mongolia University*, *Hohhot* 010021, *China*)

Abstract: Genome-wide association studies (GWAS) can systematically identify genetic varitions associated with traits and diseases. However, identifying genetic variants and risk genes significantly associated with diseases using GWAS data remain challenging. GWAS with splicing (sQTL) and expression quantitative trait loci (eQTL) data were integrated to identify key causal genes for type 2 diabetes mellitus (T2DM). Using the T2DM GWAS data and the sQTL/eQTL data derived from 49 tissues, the SMR&HEIDI methods and colocalization analysis were applied to detect potential risk causal genes. Furthermore, the differential expression analysis and functional annotation revealed their biological functions. The results show that four genes are significantly linked to T2DM and play a key role. The results offer fresh insights into T2DM's genetic mechanisms.

Key words: Type 2 diabetes mellitus; causal gene; Mendelian randomization; molecular quantitative trait loci