生物信息学分析 DOI:10.13406/j.cnki.cyxb.003389

基于转录组数据识别代谢综合征与结直肠癌的共同基因特征和分子机制

邓洁莲[1],李康杰[1],张 聪[1],张 源[1],谢 彪[1],钟晓妮[1],郑 薇[2]
（1.重庆医科大学公共卫生学院流行病与卫生统计教研室,重庆 401331;
2.重庆医科大学附属第一医院心内科,重庆 400016）

【摘 要】目的:识别代谢综合征(metabolic syndrome,MetS)与结直肠癌(colorectal cancer,CRC)共同的基因特征和生物学通路,并筛选与MetS相关的CRC预后生物标志物。**方法:**对MetS、CRC和相应的对照样本进行差异分析,以识别表达差异的基因作为与疾病相关的基因。根据这些基因的表达特征进行功能富集分析,以识别受调控的生物学功能。利用单因素Cox回归鉴定与CRC预后相关的MetS基因,并使用LASSO和多因素Cox回归分析构建CRC的预后预测模型。基于汇总数据的孟德尔随机化分析(Summary-data-based Mendelian Randomization,SMR)进一步确认MetS基因与CRC预后的因果关系。**结果:**325个基因在MetS和CRC中表达上调,281个基因表达下调。研究还发现,Apelin信号和胞吐等通路在2种疾病中被抑制,而核酸修复通路则被激活。其中,有60个与MetS和CRC共享的基因与CRC的预后相关。18个基因被用于构建预测模型,在TCGA数据库的CRC队列中,模型在1~5年的曲线下面积(area under the curve,AUC)值均大于0.75,表现出良好的预测性能。通过SMR分析进一步确定,P4HA1和LARS2与CRC预后具有因果关系。**结论:**MetS和CRC具有共同的基因特征和通路,炎症可能为2种疾病相关的基础,共享基因对CRC的预后有影响。

【关键词】代谢综合征;结直肠癌;生物信息学;预后;孟德尔随机化
【中图分类号】R735.3 **【文献标志码】**A **【收稿日期】**2023-11-28

Identification of shared genetic features and molecular mechanisms of metabolic syndrome and colorectal cancer based on transcriptomic data

Deng Jielian[1], Li Kangjie[1], Zhang Cong[1], Zhang Yuan[1], Xie Biao[1], Zhong Xiaoni[1], Zheng Wei[2]
(1. Department of Epidemiology and Health Statistics, School of Public Health, Chongqing Medical University;
2. Department of Cardiology, The First Affiliated Hospital of Chongqing Medical University)

【Abstract】*Objective*: To identify shared genetic features and biological pathways between metabolic syndrome (MetS) and colorectal cancer (CRC), and to screen for prognostic biomarkers associated with MetS in CRC. ***Methods*:** First, differential analysis was performed on MetS, CRC, and their corresponding control samples to identify genes with differential expression, which were used as disease-related genes. Subsequently, functional enrichment analysis was performed based on the expression characteristics of these genes to identify the regulated biological functions. Then, single-factor Cox regression was used to identify MetS genes associated with CRC prognosis, and LASSO and multivariate Cox regression analysis were used to construct a model for prediction of CRC prognosis. Finally, the causal relationship between MetS genes and CRC prognosis was further confirmed through Summary-data-based Mendelian Randomization. ***Results*:** A total of 325 genes were upregulated and 281 genes were downregulated in both MetS and CRC. Apelin signaling and endocytosis pathways were inhibited and the nucleotide excision repair pathway was activated in both diseases. Among them, 60 genes shared between MetS and CRC were associated with CRC prognosis. Eighteen genes were employed to construct a prediction model. In the CRC cohort of the TCGA database, the model demonstrated robust prediction performance with the area under the curve exceeding 0.75 for the 1-5 year period. Summary-data-based Mendelian Randomization analysis confirmed the causal relationship of P4HA1 and LARS2 with CRC prognosis. ***Conclusion*:** MetS and CRC share genetic features and pathways, with inflammation as a possible link between these two diseases. Shared genes can influence the prognosis of CRC.

【Key words】metabolic syndrome; colorectal cancer; bioinformatics; prognosis; Mendelian randomization

结直肠癌(colorectal cancer,CRC)是一种常见的恶性肿瘤,与多种因素有关,如年龄、家族史、饮食习惯、肥胖、缺乏运动以及长期慢性肠炎等,都可以增加患CRC的风险[1]。2020年全球新增CRC病例达到190万例,死亡病例达到94万,死亡率居各

作者介绍:邓洁莲,Email:2021120746@cqmu.edu.cn,研究方向:肿瘤风险预测。
通信作者:郑 薇,Email:vividoct@126.com。
优先出版:https://link.cnki.net/urlid/50.1046.R.20231227.1633.020(2023-12-29)

类癌症之首[2]。高发病率和高死亡率使之成为肿瘤防治领域的一大挑战。代谢综合征（metabolic syndrome，MetS）是一种聚集性疾病，包含血糖异常、血脂异常、血压异常、腹部肥胖等多种病理特征，全球近1/3的人群受其影响[3]。虽然MetS与CRC为2种不同体系的疾病，但研究表明MetS是CRC的潜在危险因素，MetS患者发生CRC的风险明显升高，而合并MetS的CRC患者更容易出现肿瘤远处转移、术后复发及不良预后等不利状况[4-5]。

随着现代生活方式和饮食结构的改变，MetS和CRC的患病率不断攀升，针对二者关联的研究逐渐显示出迫切性。目前，越来越多的证据显示MetS与CRC之间存在紧密的关联，但现有研究往往偏重于临床研究，无法深入揭示分子机制的细节。挖掘多个疾病之间共享的基因、通路、甲基化位点等，有助于揭示疾病之间的关联性、综合评估风险、确定共同生物标志物并优化治疗策略[6-7]。因此，通过生物信息学方法综合研究MetS和CRC的共同基因特征和分子机制变得尤为重要。

本研究旨在利用生物信息学方法，鉴定MetS与CRC之间共同的基因特征和生物学通路，为深入探索二者关联的分子机制实验研究提供依据和参考。此外，通过从MetS相关基因的角度筛选CRC预后生物标志物，为临床提供更加精准的疾病监测及治疗方案的选择，也为寻找CRC的治疗靶点提供新的思路。

1 资料与方法

1.1 数据收集

从GEO数据库（https://www.ncbi.nlm.nih.gov/geo）下载与CRC（GSE44076）和MetS（GSE98895）相关的转录组数据。数据包括98个CRC样本及其对应的98个正常癌旁样本，以及20个MetS样本和20个相应的正常对照样本。从TCGA数据库（https://xenabrowser.net）下载430个结肠癌样本（TCGA-COAD）和154个直肠癌样本（TCGA-READ）的转录组数据，同时还获得与这些样本相关的临床数据和预后随访信息。为进一步研究基因与表型之间的关联，本研究从eQTLGen联盟（https://www.eqtlgen.org/phase1.html）获取血液表达数量性状位点（Expression quantitative trait loci，eQTL）的统计数据，从GWAS Catalog（https://www.ebi.ac.uk/gwas/）收集与CRC总体生存率相关的GWAS数据（GCST90244553）。

1.2 差异表达分析

使用R语言limma包[8]分别对GSE98895和GSE44076数据集中病例样本和对照样本基因表达值进行差异分析。由于MetS和CRC样本分别来源于血液组织和实体组织，设置$P<0.05$且$|\log_2FC|>0$和$|\log_2FC|>0.5$分别筛选MetS和CRC的差异表达基因（differentially expressed genes，DEGs），并根据变化值的正负分为上调或下调。使用火山图进行可视化以展示病例样本中的差异表达情况和统计显著性。

1.3 通路富集分析

为了更好地理解组间基因表达差异对生物学机制的潜在影响，鉴定在MetS或CRC样本中显著富集的生物学通路和功能，使用MSigDB中包含的KEGG和GO基因集，通过clusterProfiler包[9]对差异分析的结果进行基因集富集分析（gene set enrichment analysis，GSEA）分析。

1.4 共表达网络分析

在GSE98895和GSE44076数据集上，使用R包WGCNA[10]进行加权基因共表达网络分析（weighted gene co-expression network analysis，WGCNA），以筛选共同表达的基因模块。首先，从基因表达数据中选择了具有较大中位绝对偏差的前5 000位基因作为研究对象，基于无标度拓扑准则（无标度R^2=0.9）构建邻接矩阵。随后，将邻接矩阵转化为拓扑重叠矩阵，使用平均连锁分层聚类方法对具有相似表达谱的基因模块进行聚类，将最小模块尺寸设定为30，切割高度设定为0.25识别关键模块。最后，采用Pearson相关分析计算MetS和CRC的共表达基因模块与性状的相关性。

1.5 预后分析

本研究采用单变量Cox分析筛选与CRC总生存预后的MetS相关基因。为了减少基因间的共线性影响，采用LASSO进行特征选择，并进行逐步多元Cox回归分析。通过建立预后模型，利用R包timeROC绘制受试者工作特征（receiver operating characteristic，ROC）曲线，评估这些基因对于患者预后风险的预测能力。

1.6 孟德尔随机化分析

为明确基因表达与CRC总体生存率之间的因果关系，本研究进行了基于汇总数据的孟德尔随机化分析（Summary-data-based Mendelian Randomization，SMR）[11]。该分析方法利用单核苷酸多态性作为工具变量来评估基因表达水平与CRC总体预后之间的关联。在此分析中，基因表达（汇总水平的eQTL数据）被视为暴露量，而CRC总体生存率（汇总水平的GWAS数据）作为结果变量。使用依赖仪器的异质性（Heterogeneity in Dependent Instruments，HEIDI）测试，旨在检查观察到的关联是否可能归因于连锁不平衡。如果HEIDI测试的$P>0.05$，则表明观察到的相关性不受连锁不平衡的影响。

1.7 统计学方法

使用Kaplan-Meier法和Log-rank检验比较不同风险组之间的生存曲线差异。检验水准$\alpha=0.05$。使用SMR软件的1.03版本（https://cnsgenomics.com/software/smr/#Overview）进行SMR分析和异质性检验，其余所有分析均使用R4.2.2软件进行。

2 结 果

2.1 MetS和CRC的共同差异表达基因鉴定

根据基因表达差异分析的结果显示，Mets中共有4 870

个 DEGs。其中，包括 ZNF683、SNORD13 等 2 418 个上调基因和 CCL2、LOC641 750 等 2 452 个下调基因（图 1A）。而 CRC 中共有 5 524 个 DEGs，包括 FOXQ1、CLDNI 等 2 778 个上调基因和 CA1、CLCA4 等 2 746 个下调基因（图 1B）。在 MetS 和 CRC 中，共有 325 个基因同时上调，281 个基因同时下调，即总共有 606 个基因被鉴定为 MetS 和 CRC 的共享基因（图 1C、D）。

2.2 MetS 和 CRC 的通路、功能富集情况

根据对 KEGG 通路和 GO 基因集的 GSEA 富集分析，本研究发现 MetS 和 CRC 中有多个通路和功能同时被激活或抑制。具体而言，Apelin 信号、胞吐、胰腺分泌、唾液分泌等通路在 MetS 和 CRC 中显示出明显的抑制（图 2A、B）。核受体活性、肌肉组织发育的正向调节、神经递质转运的调节、骨化的调节功能在 MetS 和 CRC 也显示出抑制趋势，而功能核苷酸切除修复在这 2 种疾病中呈现激活状态（图 2C、D）。

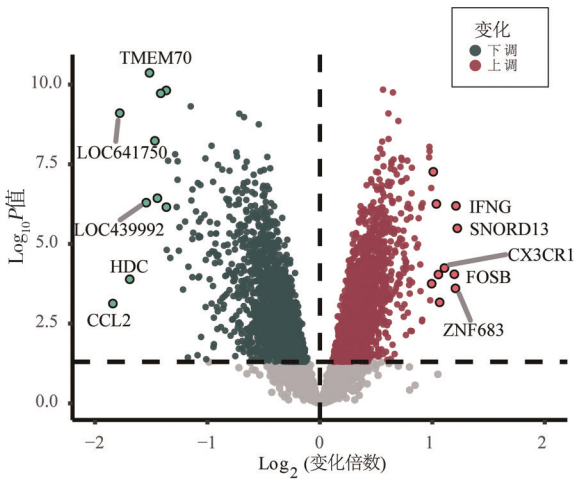
2.3 MetS 和 CRC 的加权基因共表达网络

通过对 MetS 进行 WGCNA，发现共生成 11 个基因模块（图 3A）。其中，有 5 个基因模块与 MetS 呈正相关，3 个基因模块与 MetS 呈负相关（图 3C）。同时，对 CRC 进行 WGCNA 分析，得到 7 个基因模块（图 3B）。其中，有 2 个基因模块与 CRC 呈正相关，而 3 个基因模块与 CRC 呈负相关（图 3D）。正相关的模块基因和负相关的模块基因分别鉴定为正、负相关共表达基因，本研究发现共有 68 个基因同时与 MetS 和 CRC 呈正相关，有 45 个基因为 MetS 和 CRC 的共同负相关。

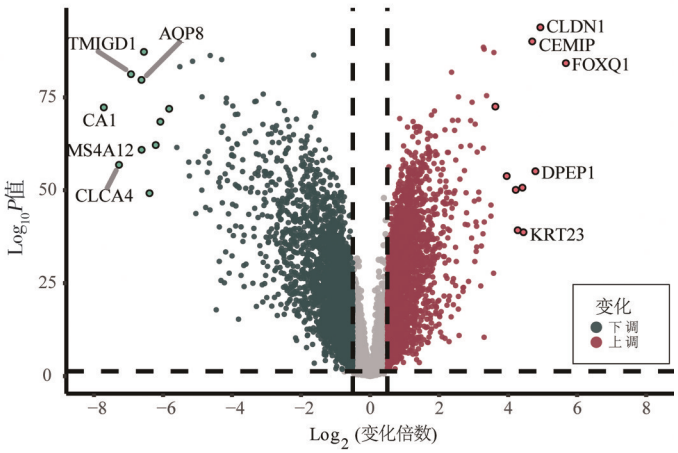
2.4 共享基因的预后价值

通过对 TCGA 数据进行单变量 Cox 回归分析，确定了与 CRC 总体生存预后相关的 60 个共享基因（图 4A）。进一步使用 LASSO 回归和多因素 Cox 分析鉴定了 18 个共享基因，用于构建 CRC 预后预测模型。该预测模型的计算公式为 Risk score=$\sum_{i=1}^{n} exp_i \times cofe_i$，其中 exp 表示基因的表达量，$cofe$ 为系数。该预测模型计算的风险评分在 1~5 年总体生存的曲线下面积（area under the curve，AUC）值均高于 0.75（图 4B），生存曲线展示了高风险组与低风险组的总体生存率之间存在明显差异（图 4C）。

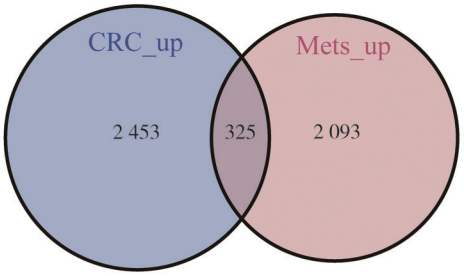
通过整合血液 eQTL 和 CRC GWAS 数据，发现 606 个共享基因中，有 466 个基因可纳入 SMR 分析。初步结果显示，其中 33 个基因的 $p_SMR<0.05$，进一步进行了 HEIDI 试验，发现其中 24 个基因的 $p_HEIDI>0.05$，即这 24 个基因通过了 SMR 和 HEIDI 试验的筛选。结合 TCGA 中单因素 Cox 分析结果，发现 P4HA1 和 LARS2 与 CRC 预后显著相关（图 5A）。在单因素 Cox 分析中，P4HA1 表达与 CRC 死亡风险呈正相关（cofe=0.20，P=0.049），在 SMR 分析中，它的表达与 CRC 风险呈正相关（b_SMR=1.22，p_SMR=0.045，p_HEIDI=0.815，图 5B），两分析的效应方向一致，均提示 P4HA1 为 CRC 预后的危险因素。在单因素 Cox 分析中，LARS2 表达与 CRC 死亡风险呈负相关（cofe=−0.549，P=0.000），在 SMR 分析中，它的表达与 CRC 风险呈负相关（b_SMR=−0.50，p_SMR=0.022，p_HEIDI=0.065，图 5D），两分析的效应方向一致，均提示 LARS2 为 CRC 预后的保护因素。



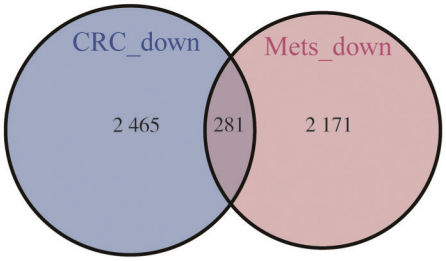
A. Mets的差异分析火山图



B. CRC的差异分析火山图



C. Mets和CRC共同上调基因



D. Mets和CRC共同下调基因

图 1 基因的差异表达分析

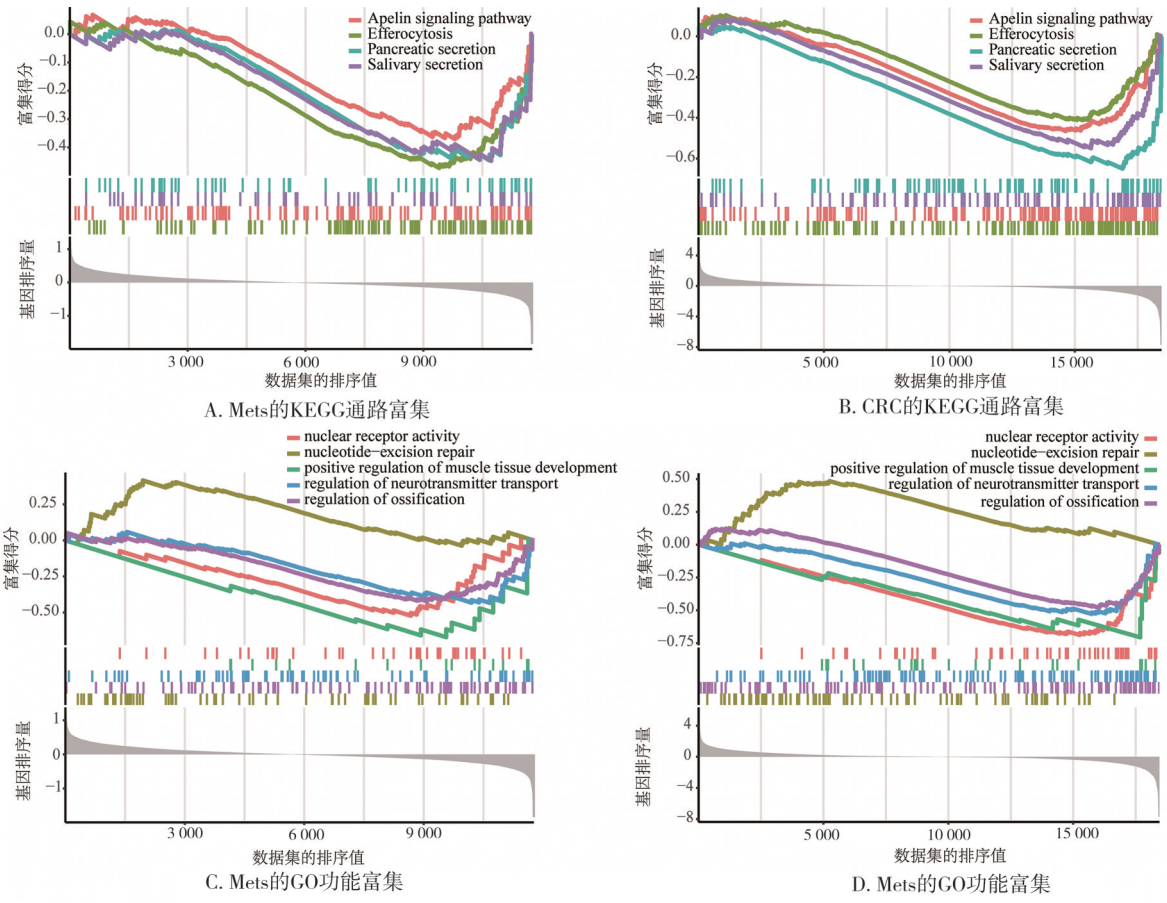


图2　明显富集的GSEA通路、功能图

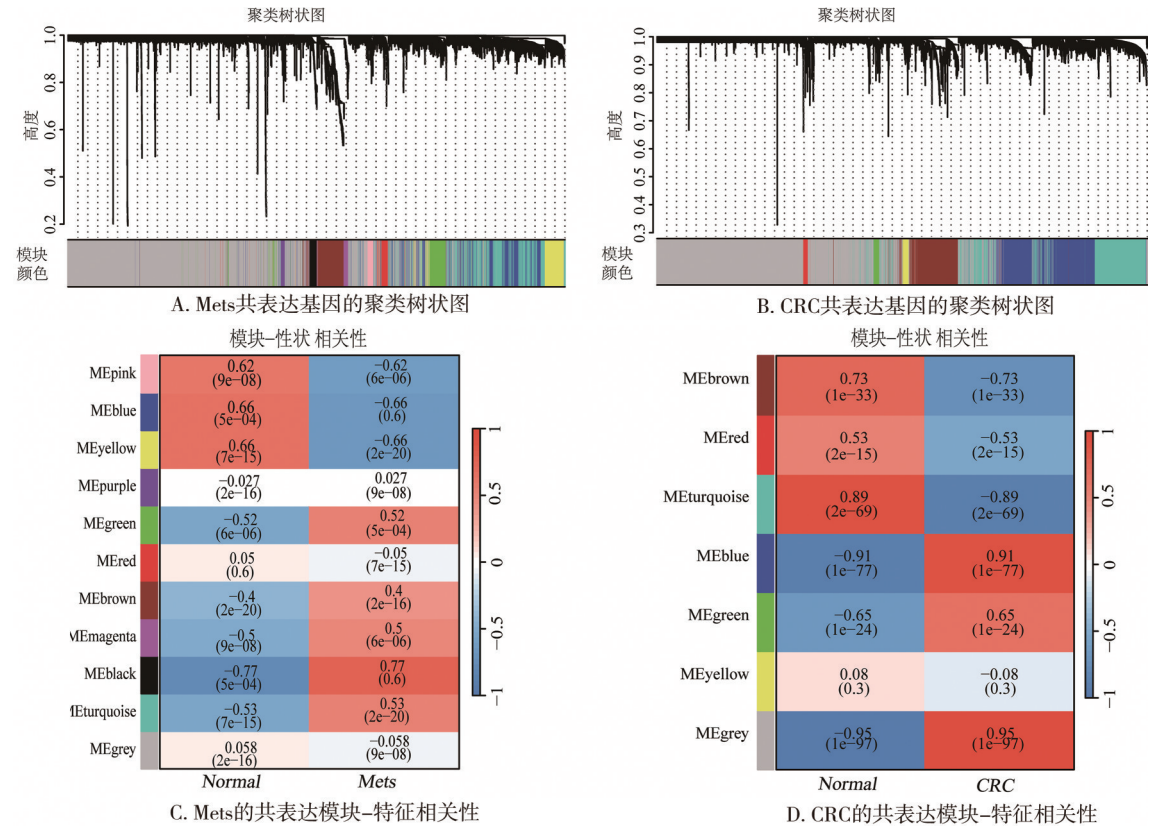
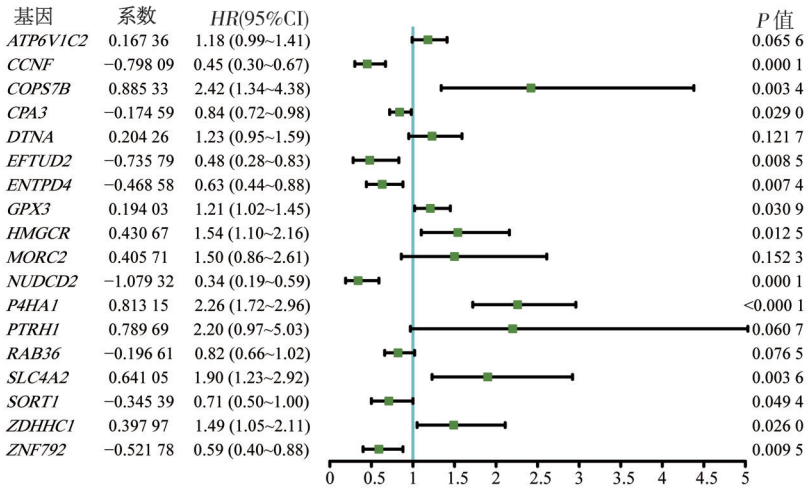
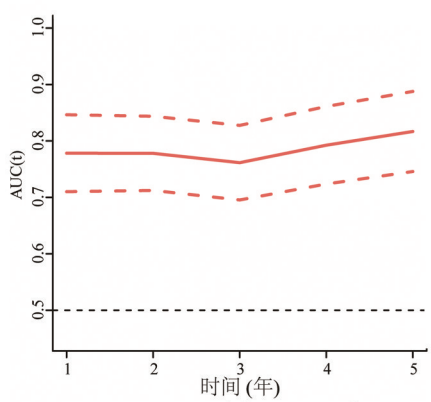


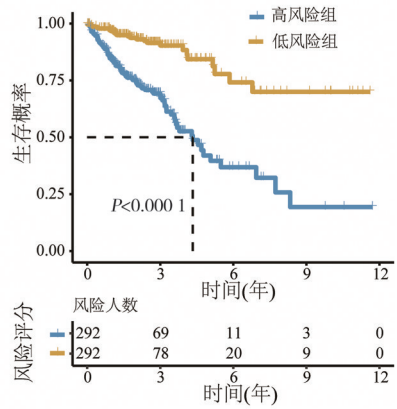
图3　基因共表达网络分析及其与疾病性状的相关性



A. 多因素Cox回归森林图

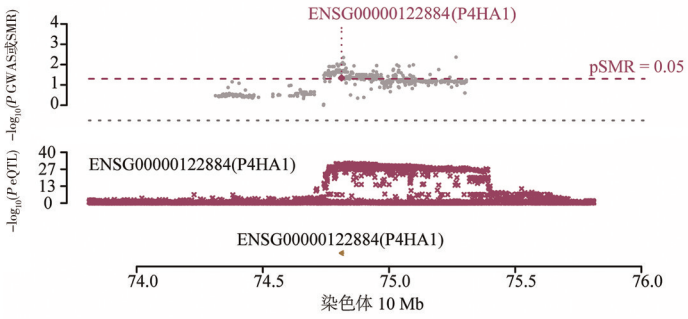


B. CRC预后风险模型的AUC值

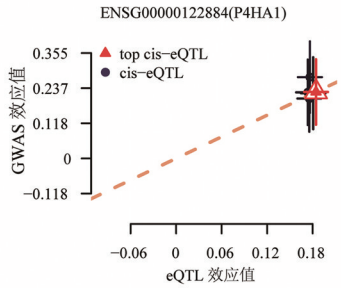


C. CRC高低风险组的KM曲线

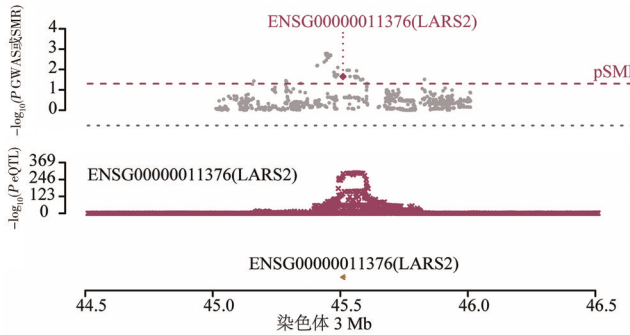
图4　MetS和CRC共同基因对CRC总体生存预后的影响



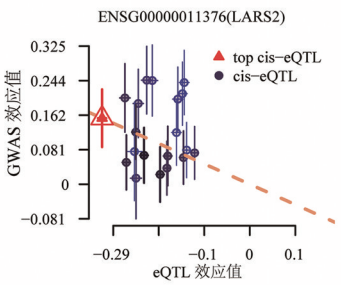
A. P4HA1相关遗传位点



B. P4HA1的SMR效应图



C. LARS2相关遗传位点



D. LARS2的SMR效应图

图5　P4HA1和LARS2与CRC总体生存预后的因果效应

3 讨 论

通过鉴定MetS与CRC之间共同的基因特征和生物学通路，本研究的发现可能为MetS和CRC之间相互作用机制进一步探索奠定基础，并影响着CRC患者的预后。

MetS和CRC共同抑制的通路，如Apelin信号和胞吐通路[12-13]，可能反映了代谢综合征与炎症之间的紧密联系。MetS通常伴随着慢性炎症状态[14]，而炎症在癌症的发展过程中被广泛认为是一个推动因素[15]。这种共同抑制表明，MetS可能引起的慢性炎症状态为结直肠癌的发展创造了有利条件，通过改变这些通路的活性来影响癌症细胞的生存和增殖。其次，核酸修复通路的激活可能与MetS引起的DNA损伤有关[16]。代谢综合征通过多种途径，如氧化应激和糖化终末产物的生成，可能导致细胞内DNA的损伤。激活核酸修复通路可能是细胞对这种损伤的生理性反应，试图维持基因组的稳定性。这一发现提供了一个新的视角，即MetS可能通过影响核酸修复过程，促进结直肠癌的发展。

MetS相关的基因进一步影响着CRC预后，P4HA1和LARS2的相反表现也需要深入思考其潜在生物学机制。P4HA1作为蛋白质羟化酶，在肿瘤中的过度表达可能与肿瘤微环境的酸化有关，进而促进肿瘤细胞的侵袭和转移[17]。相反，LARS2可能通过维持线粒体功能，抑制了癌症细胞的代谢活性和增殖[18]。这2个基因的相反表现可能是由于它们在代谢综合征与结直肠癌的不同生物学过程中扮演的角色不同。

总体而言，这些生物学机制可能是复杂而多样的，涉及多个层面的相互作用。MetS与CRC之间的交互作用可能通过影响炎症状态、DNA损伤和细胞代谢等多个生物学过程来实现。这一复杂性为未来更深入的实验研究提供了丰富的研究方法，以解开MetS与CRC关联性的生物学谜团。同时，本研究的发现，需要在以后的实验研究中进一步验证。

参 考 文 献

[1] Song MY, Bretthauer M. Interpreting epidemiologic studies of colonoscopy screening for colorectal cancer prevention: understanding the mechanisms of action is key[J]. Eur J Epidemiol, 2023, 38(9): 929-931.

[2] Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA Cancer J Clin, 2021, 71(3): 209-249.

[3] Norris P, Gow J, Arthur T, et al. Metabolic syndrome and surgical complications: a systematic review and meta-analysis of 13 million individuals[J]. Int J Surg, 2023: 201-213.

[4] Zhang CC, Zhang LT, Tian Y, et al. Association between metabolic syndrome and early-stage colorectal cancer[J]. BMC Cancer, 2023, 23(1): 1020.

[5] Han F, Wu GH, Zhang S, et al. The association of metabolic syndrome and its components with the incidence and survival of colorectal cancer: a systematic review and meta-analysis[J]. Int J Biol Sci, 2021, 17(2): 487-497.

[6] Rahman MR, Islam T, Shahjaman M, et al. Discovering common pathogenetic processes between COVID-19 and diabetes mellitus by differential gene expression pattern analysis[J]. Brief Bioinform, 2021, 22(6): bbab262.

[7] Gallon J, Coto-Llerena M, Ercan CE, et al. Epigenetic priming in chronic liver disease impacts the transcriptional and genetic landscapes of hepatocellular carcinoma[J]. Mol Oncol, 2022, 16(3): 665-682.

[8] Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies[J]. Nucleic Acids Res, 2015, 43(7): e47.

[9] Wu TZ, Hu EQ, Xu SB, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data[J]. Innovation (Camb), 2021, 2(3): 100141.

[10] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis[J]. BMC Bioinformatics, 2008, 9: 559.

[11] Zhu ZH, Zhang FT, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets[J]. Nat Genet, 2016, 48(5): 481-487.

[12] Sawane M, Kidoya H, Muramatsu F, et al. Apelin attenuates UVB-induced edema and inflammation by promoting vessel function[J]. Am J Pathol, 2011, 179(6): 2691-2697.

[13] Weimershaus M, Carvalho C, Rignault R, et al. Mast cell-mediated inflammation relies on insulin-regulated aminopeptidase controlling cytokine export from the Golgi[J]. J Allergy Clin Immunol, 2023, 151(6): 1595-1608.

[14] Šebeková K, Staruchová M, Mišľanová C, et al. Association of inflammatory and oxidative status markers with metabolic syndrome and its components in 40-to-45-year-old females: a cross-sectional study[J]. Antioxidants, 2023, 12(6): 1221.

[15] Greten FR, Grivennikov SI. Inflammation and cancer: triggers, mechanisms, and consequences[J]. Immunity, 2019, 51(1): 27-41.

[16] Milanese C, Bombardieri CR, Sepe S, et al. DNA damage and transcription stress cause ATP-mediated redesign of metabolism and potentiation of anti-oxidant buffering[J]. Nat Commun, 2019, 10(1): 4887.

[17] Liu Y, Gu YJ, Ng S, et al. Circulating levels of hydroxylated bradykinin function as an indicator of tissue HIF-1α expression[J]. Sci Bull, 2020, 65(18): 1570-1579.

[18] Li ZJ, Bao HY. Comparative transcriptome and proteome analysis explores the antitumor key regulators of ergosterone in H22 tumor-bearing mice[J]. Eur J Pharmacol, 2023, 953: 175831.

（责任编辑：周一青）