·论　　著·

基于生物信息分析的结直肠癌表达和预后的关键基因筛选

陈晓玲[1]，杨　娟[2]，吕敏敏[3]，贾亦真[4△]

（1. 香港大学深圳医院感染性疾病医学部，广东 深圳 518000；2. 深圳市坪山区妇幼保健院检验科，广东 深圳 518000；3. 香港大学深圳医院中心实验室，广东 深圳 518000；4. 香港大学深圳医院科研部临床研究管理办公室，广东 深圳 518000）

［摘　要］ **目的**　识别结直肠癌中的差异表达基因（DEGs），探索结直肠癌的关键通路和基因。**方法**　选取来自基因表达综合数据集的基因表达谱 GSE211496、GSE6988 和 GSE29900 数据集，使用 GEO2R 分析工具进行分析并下载相关数据，通过在线数据库 miRDB 和 TargetScan 对 GSE29900 数据集的差异 miRNAs 进行靶基因预测，然后使用韦恩图取 3 个数据库的 DEGs 交集，使用 DAVID 数据库工具进行 GO 和 KEGG 通路富集分析，接着使用 PPI 进行网络构建并由 Cytoscape 软件进行可视化，使用 TCGA 数据库验证 Hub 基因表达，使用 pROC 包对与 TCGA 数据库表达一致的 Hub 基因进行 ROC 曲线分析，最后，利用 Kaplan-Meier 绘图仪在线工具对结直肠癌患者进行预后分析。**结果**　筛选出 GSE211496 数据集 2 570 个 DEGs（p. adj＜0. 01 且 $|\log_2 FC| \geq 1$），GSE6988 数据集 406 个 DEGs（p. adj＜0. 01 且 $|\log_2 FC| \geq 1$）和 GSE29900 数据集 99 个差异表达 miRNA（p. adj＜0. 01 且 $|\log_2 FC| \geq 1$），预测出差异表达 miRNAs 的靶基因 14 938 个，将靶基因与 DEGs 重叠共获得 30 个目标基因。KEGG 通路结果显示，目标基因主要富集于血管平滑肌收缩和矿物吸收通路。通过连接度从 PPI 网络中筛选出前 10 个 Hub 基因；Hub 基因经 TCGA 数据库验证，发现 MYL9、ACTG2、AGT 和 PDGFRA 与 GSE211496 数据集表达一致。分析这 4 个 Hub 基因对结直肠癌的诊断情况发现，基因 AGT（AUC＝0. 901，95%*CI* 0. 868～0. 933）与预测结直肠癌的发生呈正相关，基因 MYL9（AUC＝0. 820，95%*CI* 0. 757～0. 884）、ACTG2（AUC＝0. 855，95%*CI* 0. 802～0. 908）和 PDGFRA（AUC＝0. 815，95%*CI* 0. 772～0. 858）与预测结直肠癌的发生呈负相关。基因 MYL9、ACTG2 和 PDGFRA 对结直肠癌诊断均有一定准确性，基因 AGT 对结直肠癌诊断具有较高准确性。Kaplan-Meier 生存分析发现，PDGFRA、ACTG2 和 MYL9 低表达均显示患者预后较好，差异均有统计学意义（P＜0. 05）。**结论**　该研究通过生物信息学分析筛选并鉴定出 4 个基因是结直肠癌中的枢纽基因，这些基因包括 PDGFRA、ACTG2、MYL9 和 AGT，这将为结直肠癌研究提供一些新方向。

［关键词］ 结直肠癌；　生物信息学；　差异表达基因；　ROC 曲线分析；　生存分析

DOI：10. 3969/j. issn. 1009-5519. 2024. 07. 006　　**中图法分类号**：R318. 04

文章编号：1009-5519（2024）07-1109-09　　**文献标识码**：A

Key gene screening for colorectal cancer expression and prognosis based on bioinformatics analysis

CHEN Xiaoling[1], *YANG Juan*[2], *LV Minmin*[3], *JIA Yizhen*[4△]

（1. *Department of Infectious Diseases, Shenzhen Hospital of University of Hong Kong, Shenzhen, Guangdong* 518000, *China*; 2. *Clinical Laboratory, Pingshan District Maternal and Child Health Hospital, Shenzhen, Guangdong* 518000, *China*; 3. *Central Laboratory, Shenzhen Hospital of University of Hong Kong, Shenzhen, Guangdong* 518000, *China*; 4. *Department of Research Management Office, Department of Scientific Research, Shenzhen Hospital of University of Hong Kong, Shenzhen, Guangdong* 518000, *China*）

［**Abstract**］ **Objective**　To identify differentially expressed genes（DEGs）in colorectal cancer and explore key pathways and genes in colorectal cancer due to the relatively high incidence of colorectal cancer. **Methods**　The gene expression profiles GSE211496, GSE6988 and GSE29900 data sets from the gene expression comprehensive data set were selected, and the GEO2R analysis tool was used to analyze and download the relevant data. The online database miRDB and TargetScan were used to predict the target genes of the differential miRNAs in the GSE29900 data set. Then, the DEGs intersection of the three databases was taken by

作者简介：陈晓玲（1989－），本科，主管检验师，主要从事肿瘤、免疫、感染性疾病等相关研究。　△　**通信作者**，E-mail：jiayz@hku-szh. org。

Wayne diagram, and the DAVID database tool was used for Gene Ontology(GO) and Kyoto Encyclopedia of Genes and Genomes(KEGG) pathway enrichment analysis. Then Protein-Protein Interactions(PPI) protein interaction tool was used for network construction and visualization by Cytoscape software. Hub gene expression was verified by The Cancer Genome Atlas(TCGA) database. Receiver operating characteristic(ROC) curve analysis was performed on Hub genes consistent with TCGA database expression using pROC package. Finally, Kaplan-Meier plotter online tool detection was used to analyze the prognosis of colorectal cancer patients. **Results** A total of 2 570 DEGs(p. adj<0. 01 and $|\log_2 FC| \geq 1$) in GSE211496 dataset, 406 DEGs(p. adj<0. 01 and $|\log_2 FC| \geq 1$) in GSE6988 dataset and 99 differentially expressed miRNAs(p. adj<0. 01 and $|\log_2 FC| \geq 1$) in GSE29900 dataset were screened out, and 14 938 target genes of differentially expressed miRNAs were predicted. A total of 30 target genes were obtained by overlapping the target genes with DEGs. The results of KEGG pathway showed that the target genes were mainly enriched in vascular smooth muscle contraction and mineral absorption pathways. The top 10 Hub genes were screened from the PPI network by connectivity. The TCGA database of Hub genes verified that MYL9, ACTG2, AGT and PDGFRA were consistent with the expression of GSE211496 data set. Analysis of the four Hub genes for the diagnosis of colorectal cancer showed that gene AGT(AUC=0. 901, 95%*CI* 0. 868-0. 933) was positively correlated with the prediction of colorectal cancer, while gene MYL9(AUC=0. 820, 95%*CI* 0. 757-0. 884), gene ACTG2(AUC=0. 855, 95%*CI* 0. 802-0. 908) and gene PDGFRA(AUC=0. 815, 95%*CI* 0. 772-0. 858) were negatively correlated with the prediction of colorectal cancer. Gene MYL9, ACTG2 and PDGFRA had certain accuracy in the diagnosis of colorectal cancer, and gene AGT had high accuracy in the diagnosis of colorectal cancer. Kaplan-Meier survival analysis showed that low expression of PDGFRA, ACTG2 and MYL9 indicated better prognosis, and the differences were statistically significant($P<0. 05$). **Conclusion** In this study, four genes were screened and identified as hub genes in colorectal cancer by bioinformatics analysis. These genes include PDGFRA, ACTG2, MYL9 and AGT, which will provide some new directions for colorectal cancer research.

[**Key words**] Colorectal cancer; Bioinformatics; Differentially expressed genes; Receiver operating characteristic curve analysis; Survival analysis

结直肠癌是常见的消化道恶性肿瘤之一，据有关研究报道结果显示，结直肠癌的恶性程度居世界第 3 位，死亡率居第 2 位[1-3]。根据 2020 年我国癌症统计报告结果显示，结直肠癌的发病率和死亡率均呈上升趋势，且有年轻化表现[4]。本研究通过选取 GSE211496、GSE6988 和 GSE29900 这 3 个数据集，使用 GEO2R 分析工具进行分析并下载相关数据，通过在线数据库 miRDB 和 TargetScan 对 GSE29900 数据集的差异 miRNA 进行靶基因预测，然后使用韦恩图取 3 个数据库的差异表达基因(DEGs)交集，使用数据库注释、可视化和集成发现(DAVID)工具进行基因本体(GO)功能和京都基因与基因组百科全书(KEGG)通路富集分析，接着利用蛋白质相互作用(PPI)进行网络构建并由 Cytoscape 软件进行可视化，探索结直肠癌的关键通路和基因，接着用 TCGA 数据库验证 Hub 基因表达，对与 TCGA 数据库表达一致的 Hub 基因进行受试者工作特征(ROC)曲线分析，最后，利用 Kaplan-Meier 绘图仪在线工具检测对结直肠癌患者进行预后分析。这种研究分析方法与多篇已报道的文章相类似[5-8]。本研究旨在识别结直肠癌中的 DEGs，探索结直肠癌的关键通路和基因，进一步阐明结直肠癌发生发展的分子机制，为后续寻找临床诊断标志物及治疗靶点提供重要线索。

1 材料与方法

1.1 材料

1.1.1 流程图 见图 1。

1.1.2 研究材料 GEO 数据库(http://www.ncbi.nlm.nih.gov/geo/)[9] 选取 GSE211496、GSE6988 和 GSE29900 数据集，全部来源为人类。其中 GSE211496 数据集正常结直肠上皮细胞 4 例，结直肠癌细胞 4 例。GSE6988 数据集中正常结直肠上皮细胞 28 例，结直肠癌细胞 82 例。GSE29900 数据集正常结直肠上皮细胞 1 例，结直肠癌细胞 9 例。其中 GSE211496 和 GSE6988 数据集的差异基因类型是 DEGs，而 GSE2990 数据集的差异基因类型为 miRNAs。

1.2 方法

1.2.1 筛选 DEGs 和差异表达 miRNA 使用 GEO2R 分析工具(https://www.ncbi.nlm.nih.gov/geo/geo2r/)分别对 GSE211496、GSE6988 和 GSE29900 数据集进行分析并下载相关数据，以 p. adj<0. 01 且 $|\log_2 FC| \geq 1$ 为标准筛选 GSE211496、GSE6988 数据集的 DEGs 和 GSE29900 数据集的差异表达 miR-

NAs。使用 GEO2R 工具进行可视化分析，并生成火山图。见图 2。

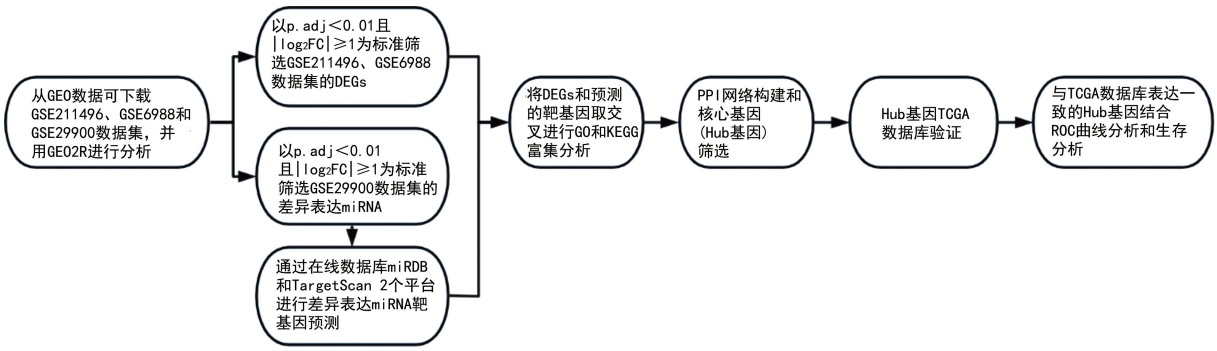


图 1　流程图

1.2.2　miRNA 靶基因预测　通过在线数据库 miRDB[10]（http://www.mirdb.org/mirdb/index.html）和 TargetScan[11]（https://www.targetscan.org/vert_80/）2 个平台进行差异表达 miRNA 靶基因预测，采用 Bioinformatics Evolutionary Genomics（http://bioinformatics.psb.ugent.be/webtools/Venn/）将 2 个平台的预测结果与前面筛选的 DEGs 取交集，得到目标基因。

1.2.3　GO 和 KEGG 通路富集分析　使用在线工具 DAVID[12]（https://david.ncifcrf.gov/home.jsp）对筛选的目标基因进行 GO 功能和 KEGG 通路富集分析，GO 功能分析包括生物学过程（BP）、细胞组分（CC）和分子功能（MF），以 $P<0.05$ 为有效。

1.2.4　PPI 网络构建和核心基因（Hub 基因）筛选　通过在线分析工具 STRING[13]（https://string-db.org/）预测并构建目标基因的 PPI 网络，将本研究筛选出的所有 DEGs 导入 STRING，通过 STRING 分析工具可以进一步探寻这些 DEGs 之间潜在的联系[14]。筛选条件设置为：置信度≥0.15，互作最大值＝0。之后把 STRING 的计算结果导入 Cytoscape3.9.0[15]，挖掘 PPI 网络中连接最为紧密的集簇[16]，应用的是分子复合物检测（MCODE）插件，筛选设置参数设置为默认参数。此外，以 $\log_2$FC 正值的为高表达组，负值的为低表达组，表达值使用数据集 GSE211496，应用 cytoHubba 插件筛选出 PPI 网络中处于关键位置的前 10 个基因。

1.2.5　Hub 基因 TCGA 数据库验证和 ROC 诊断曲线分析　从 TCGA 数据库（https://portal.gdc.cancer.gov）下载并整理 TCGA-COAD 和 TCGA-READ 项目 STAR 流程的 RNAseq 数据，提取 TPM 格式的数据，使用 pROC 包对数据进行 ROC 分析，结果用 ggplot2 进行可视化。ROC 诊断曲线分析，在曲线下面积（AUC）＞0.5 的情况下，AUC 越接近于 1，说明该变量在预测结局上诊断效果越好。0.5＜AUC＜0.7 时有较低准确性，0.7≤AUC≤0.9 时有一定准确性，AUC＞0.9 时有较高准确性。AUC＝0.5 时，说明该变量不起作用，无诊断价值。

1.2.6　生存分析　使用 Kaplan-Meier Plotter 数据库[17]（http://kmplot.com/analysis/）评估 Top10 Hub 基因与直肠腺癌患者总体生存率的关系，绘制 Kaplan-Meier 生存曲线。$P<0.05$ 为差异有统计学意义。

2　结　果

2.1　结直肠癌细胞和正常结直肠上皮细胞中 DEGs 和差异表达 miRNAs　GSE211496 数据集 gene ID 的总数为 175 143 个，其中，满足 p.adj＜0.01 且 $|\log_2FC|\geq1$ 阈值的 ID 有 2 570 个，在这阈值下，高表达（$\log_2$FC 为正）的数目有 1 423 个，低表达（$\log_2$FC 为负）的数目有 1 147 个；GSE6988 数据集 ID 的总数为 15 539 个，其中，满足 p.adj＜0.01 且 $|\log_2FC|\geq1$ 阈值的 ID 有 406 个，在这阈值下，高表达（$\log_2$FC 为正）的数目有 70 个，低表达（$\log_2$FC 为负）的数目有 336 个；GSE29900 数据集 ID 的总数为 3 069 个，其中，满足 p.adj＜0.01 且 $|\log_2FC|\geq1$ 阈值的 ID 有 99 个，在这阈值下，高表达（$\log_2$FC 为正）的数目有 4 个，低表达（$\log_2$FC 为负）的数目有 95 个。

表 1　GSE29900 数据集筛选的差异表达 miRNAs

miRNA_ID	$\log_2$FC	P	p.adj
hsa-miR-145	−11.067 644	0.000 076 04	0.007 37
hsa-miR-199a	−9.410 823	0.000 145 52	0.007 57
hsa-miR-215	−7.526 101	0.000 171 08	0.007 65
hsa-miR-23b	−5.558 338	0.000 126 99	0.007 43
hsa-miR-557	−5.150 836	0.000 053 96	0.007 37
hsa-miR-223	−5.086 244	0.000 004 77	0.005 46
hsa-miR-638	−4.095 176	0.000 547 20	0.007 37
hsa-miR-572	−4.028 467	0.000 068 68	0.007 37
hsa-miR-422b	−3.681 826	0.000 082 13	0.007 37
hsa-miR-671	−3.515 486	0.000 301 54	0.009 44
hsa-miR-18a*	3.586 709	0.000 292 54	0.009 26
hsa-miR-301	5.858 650	0.000 079 36	0.007 37
hsa-miR-183	6.888 223	0.000 073 03	0.007 37

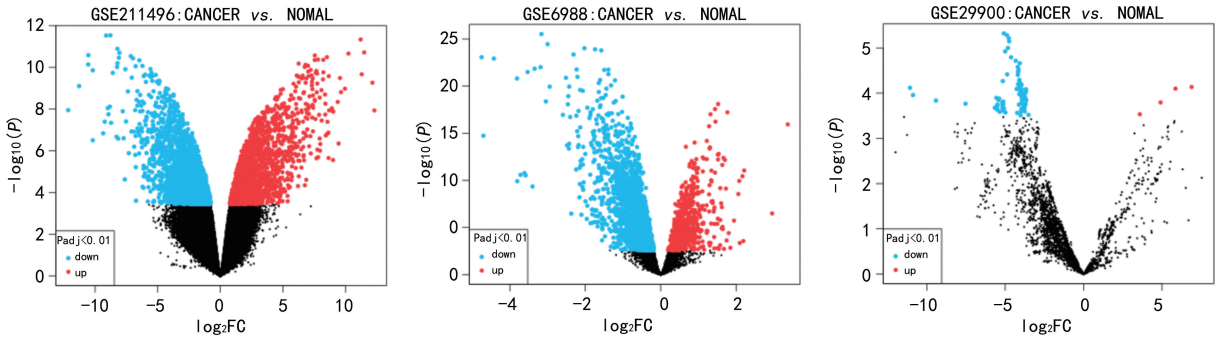


图 2　GSE211496、GSE6988 数据集中 DEGs 和 GSE29900 数据集中差异表达 miRNAs 火山图

2.2　差异表达 miRNAs 靶基因的预测　对 GSE29900 数据集取 $|\log_2 FC|>1$ & p.adj<0.01 阈值，并删除重复项后，共筛选出 14 个 miRNAs，见表 1。采用 TargetScan 和 miRDB 预测差异表达 miRNAs 的靶基因，并对 2 个数据库的结果进行重叠，删除重复项，共预测出 14 938 个靶基因。与前面 2 个数据集的差异 DEGs 取交集得到 30 个目标基因，见图 3；30 个目标基因的具体信息，见表 2。

表 2　30 个目标基因的具体信息

目标基因	$\log_2 FC$	*P*	p.adj	基因描述
REG1A	−6.61	8.40E−08	3.70E−05	再生胰岛衍生的 1α(胰腺结石蛋白、胰腺线蛋白)
MMP1	−5.42	1.22E−06	2.29E−04	基质金属蛋白酶 1(间质胶原酶)
HSPB8	−5.14	1.83E−08	1.36E−05	蛋白激酶 H11、小应激蛋白样蛋白 HSP22
MYL9	−3.99	1.14E−07	4.46E−05	肌球蛋白调节轻链 2、平滑肌亚型
SLC26A2	−3.08	1.10E−04	4.87E−03	溶质载体家族 26(硫酸盐转运体)，成员 2
ACTG2	−3.03	3.57E−05	2.39E−03	肌动蛋白、γ-2、平滑肌、肠
EDN3	−3.02	3.97E−05	2.57E−03	内皮素 3
SEMA3C	−3.01	1.91E−05	1.60E−03	Sema 结构域、免疫球蛋白结构域(Ig)、短碱性结构域、分泌型
ANLN	−2.27	1.13E−04	4.96E−03	Anillin(果蝇 Scraps 同源物)、肌动蛋白结合蛋白
FNBP1	−2.18	3.06E−04	9.15E−03	KIAA0554 蛋白
KALRN	−2.06	1.07E−04	4.79E−03	具有 Dbl-和 pleckstrin 同源结构域的丝氨酸/苏氨酸激酶
PDGFRA	−2.03	1.43E−05	1.31E−03	血小板衍生生长因子受体、α 多肽
MALL	−1.90	1.05E−06	2.06E−04	BENE 蛋白
IDH2	1.44	1.91E−05	1.60E−03	异柠檬酸脱氢酶 2($NADP^+$)、线粒体
AZGP1	1.83	9.36E−05	4.38E−03	α-2-糖蛋白 1、锌
PMAIP1	1.88	9.82E−06	1.02E−03	佛波醇-12-肉豆蔻酸-13-乙酸酯诱导蛋白 1
ST6GALNAC6	1.88	3.03E−04	9.11E−03	CMP NeuAC:(β)-N-乙酰氨基半乳糖(α)2,6-唾液酰基转移酶成员Ⅵ
ASRGL1	2.26	4.68E−05	2.83E−03	假定蛋白 FLJ22316
TSPAN1	2.48	1.99E−05	1.63E−03	四旋蛋白 1
SLC22A18AS	2.96	2.76E−07	7.99E−05	溶质载体家族 22(有机阳离子转运体)、成员 1 - 像反义
KLRC1	4.14	2.27E−07	7.06E−05	杀伤细胞凝集素样受体亚家族 C、成员 1
CA2	4.20	3.25E−05	2.26E−03	碳酸酐酶Ⅱ
PMP22	4.20	3.71E−08	2.17E−05	外周髓鞘蛋白 22
PAPSS2	4.47	1.52E−04	5.94E−03	3′-磷酸腺苷 5′-磷酸硫酸合酶 2
PRKAR2B	4.75	2.77E−05	2.04E−03	蛋白激酶、cAMP 依赖性、调节性、Ⅱ型
MT1G	6.10	1.75E−06	2.91E−04	金属硫蛋白 1G
MT1A	7.68	2.28E−06	3.53E−04	ESTs、与 SMHU1B 金属硫蛋白 1B[H.智人]高度相似
AGT	7.97	9.89E−07	1.98E−04	丝氨酸(或半胱氨酸)蛋白酶抑制剂、A 分支(α-1 抗蛋白酶，抗胰蛋白酶)、成员 8
CDH3	8.64	1.34E−07	4.88E−05	Cadherin 3，1 型、胎盘钙黏素(胎盘型)
MT1E	1.13E+01	2.13E−10	9.39E−07	金属硫蛋白 1E(功能性)

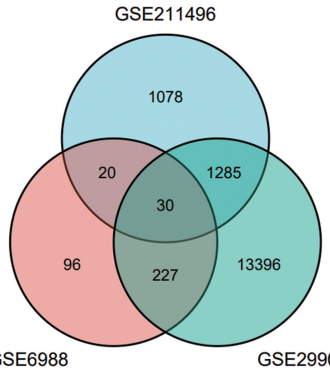
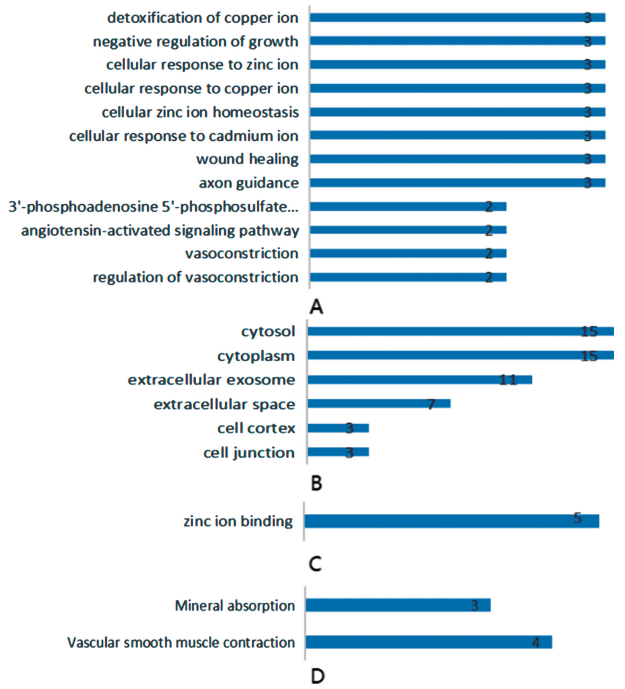


图 3　DEGs 与预测的靶基因交集的韦恩图

2.3　目标基因的 GO 功能和 KEGG 通路富集分析　在满足 p. adj＜0. 05 条件下，BP 共有 12 条，CC 共有 6 条，MF 共有 1 条，KEGG 共有 2 条。GO 功能富集分析显示：目标基因富集于规律血管收缩、血管紧张素激活信号通路、3'-磷酸腺苷 5'-磷酸硫酸盐、轴突导引、伤口愈合、细胞对镉离子的反应、细胞锌离子稳态、细胞对铜离子的反应、细胞对锌离子的反应、生长负调节、铜离子解毒等 BP；富集于细胞连接、细胞皮层、细胞外液、胞外体、细胞质、胞质溶胶等 CC；富集于锌离子结合 MF。KEGG 通路富集分析显示：目标基因参与血管平滑肌收缩和矿物吸收通路。见表 3、图 4。



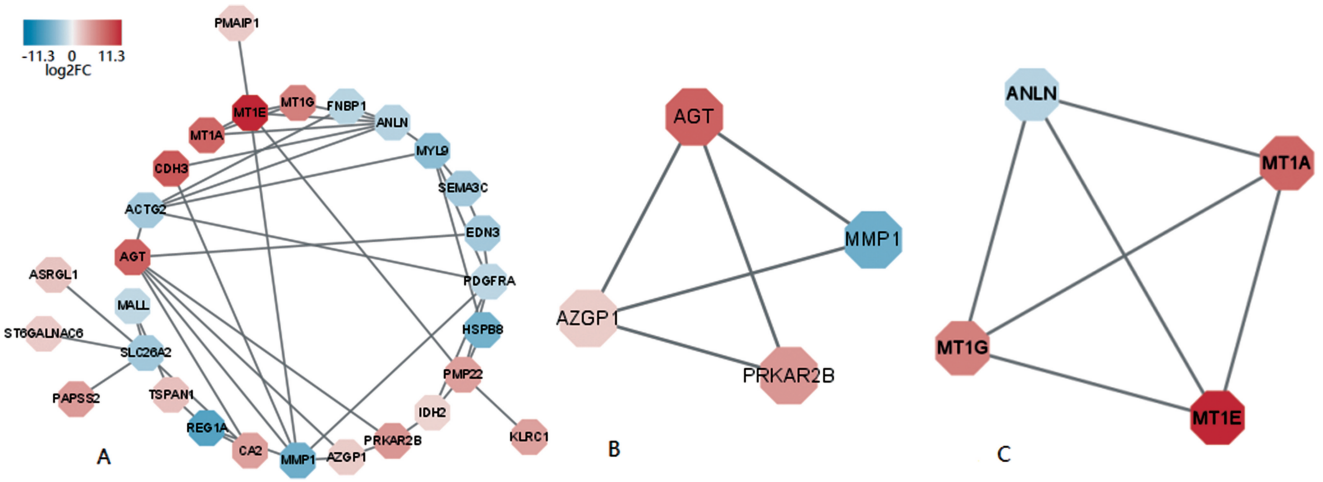
注：A. 生物过程；B. 细胞组成；C. 分子功能；D：目标基因的 KEGG 信号通路分析。

图 4　目标基因的 GO 功能(A、B、C)和 KEGG 通路(D)富集分析

表 3　目标基因的 GO 功能和 KEGG 通路富集分析

种类	名称	数量(个)	*P*
BP	GO:0071280～细胞对铜离子的反应	3	8.08E－04
BP	GO:0071294～细胞对锌离子的反应	3	5.99E－04
BP	GO:0019229～规律血管收缩	2	0.028 697 561
BP	GO:0042310～血管收缩	2	0.028 697 561
BP	GO:0038166～血管紧张素激活信号通路	2	0.015 883 324
BP	GO:0050428～3′-磷酸腺苷、5′-磷酸硫酸盐	2	0.008 694 024
CC	GO:0005737～细胞质	15	0.008 502 231
CC	GO:0005829～细胞溶质	15	0.007 250 922
CC	GO:0070062～细胞外分泌体	11	4.55E－04
CC	GO:0005615～细胞外液	7	0.043 631 385
CC	GO:0030054～细胞连接	3	0.039 844 794
CC	GO:0005938～细胞皮层	3	0.021 036 092
MF	GO:0008270～锌离子结合	5	0.038 887 132
KEGG	hsa04270：血管平滑肌收缩	4	0.004 028 809
KEGG	hsa04978：矿物质吸收	3	0.009 286 634

2.4　DEGs 的 PPI 网络分析结果　通过 STRING 得到的 PPI 网络有 28 个节点蛋白，将 PPI 网络导入 Cytoscape 本地版，根据 $\log_2$FC 大小美化 PPI 网络(图 5A)。通过 MCODE 插件分析发现 2 组集簇，其中 1 个集簇包含 4 个节点(node)与 5 条连线(edge)(图 5B)，另一个集簇包含 4 个节点(node)与 6 条连线(edge)(图 5C)。通过 cytoHubba 插件按照 MCC 为标准，最终筛选出处于关键位置的 10 个基因(图 6)，分别为 PMP22、MT1E、MT1G、MT1A、ANLN、MYL9、ACTG2、AGT、MMP1 和 PDGFRA。



注：A. PPI 网络分析结果；B. 聚类模块由 MCODE 获得。

图 5　PPI 网络分析结果和集簇模块构建

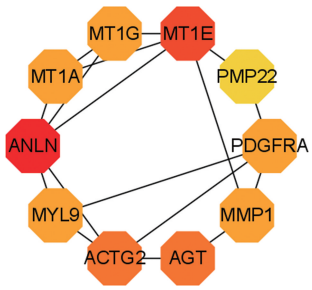
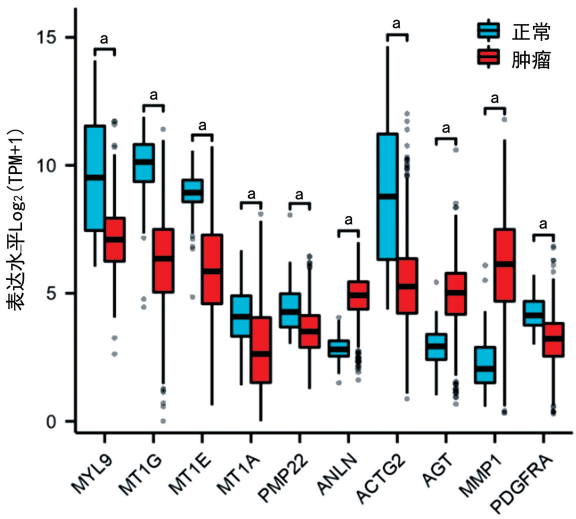


图 6　cytoHubba 插件筛选出 PPI 网络中处于关键位置的前 10 个基因

2.5　Hub 基因 TCGA 数据库验证　将筛选的 Hub 基因 PMP22、MT1E、MT1G、MT1A、ANLN、MYL9、ACTG2、AGT、MMP1 和 PDGFRA 用 TCGA 数据库进行验证，发现 MYL9、ACTG2、AGT 和 PDGFRA 与 GSE211496 数据集表达一致。见图 7。



注：与正常比较，$^{a}P<0.05$。

图 7　Hub 基因在 TCGA 数据库的表达水平图

2.6　与 TCGA 数据库表达一致的 Hub 基因结合 ROC 曲线分析　将筛选的与 TCGA 数据库表达一致的 Hub 基因 MYL9、ACTG2、AGT 和 PDGFRA 结合 ROC 曲线，分析 Hub 基因对结直肠癌的诊断情况，AGT（AUC=0.901，95%*CI* 0.868～0.933）与预测结局呈正相关，MYL9（AUC=0.820，95% *CI* 0.757～0.884）、ACTG2（AUC=0.855，95% *CI* 0.802～0.908）和 PDGFRA（AUC=0.815，95% *CI* 0.772～0.858）预测结直肠癌的结局为反向相关。基因 MYL9、ACTG2 和 PDGFRA 对结直肠癌诊断均有一定准确性，基因 AGT 对结直肠癌诊断具有较高准确性。见图 8。

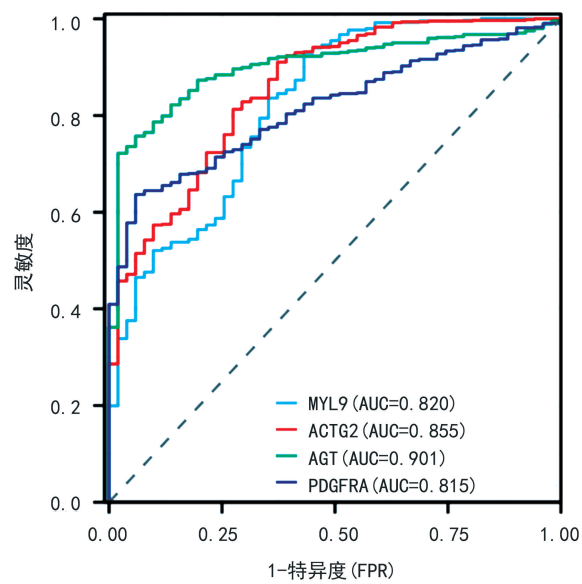


图 8　与 TCGA 数据库表达一致的 Hub 基因结合 ROC 曲线分析图

2.7　与 TCGA 数据库表达一致的 Hub 基因结合生存曲线分析　将筛选的与 TCGA 数据库表达一致的 Hub 基因 MYL9、ACTG2、AGT 和 PDGFRA 结合患者生存曲线，分析 Hub 基因对患者总体生存率的影响。PDGFRA、ACTG2 和 MYL9 差异均有统计学意义（$P<0.05$），其中 PDGFRA、ACTG2 和 MYL9 低表达均显示患者预后较好。见图 9。

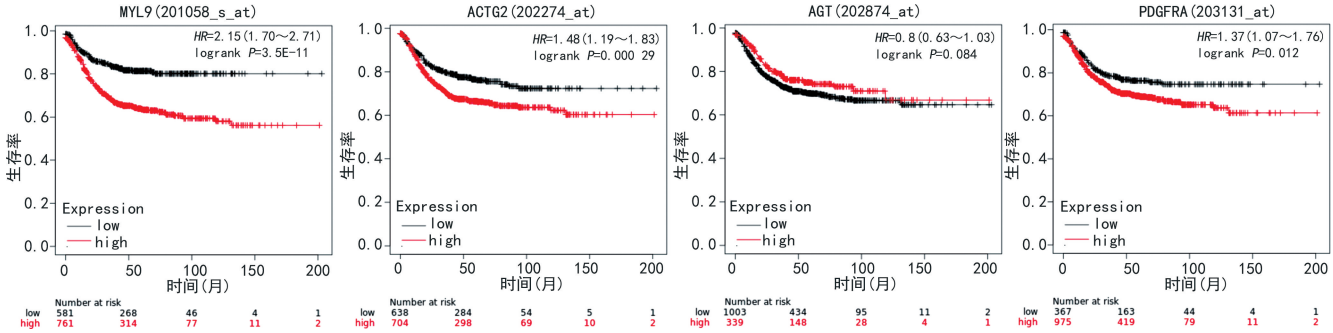


图 9　与 TCGA 数据库表达一致的 Hub 基因表达与患者总体生存率的关系

3　讨　论

大量文献研究了与结直肠癌相关的不同表达基因[18-20]。然而，结直肠癌的发病机制尚不清楚，所以有必要利用最新的数据继续分析结直肠癌中不同表达的基因。

本研究中，选取 GSE211496、GSE6988 和 GSE29900 数据集，包含正常结直肠上皮细胞 38 例，结直肠癌细胞 90 例，从 GEO 数据库下载，提取目标基因。GSE211496 数据集高表达的数目有 1 423 个，低表达的数目有 1 147 个；GSE6988 数据集高表达的数目有 70 个，低表达的数目有 336 个；GSE29900 数据集为 miRNAs，高表达的数目有 4 个，低表达的数目有 95 个，筛选此数据集的差异表达 miRNAs，并用 miRDB 和 TargetScan 平台进行差异基因靶基因预测，共预测出 14 938 个靶基因后，将预测出的靶基因与前面 2 个数据集的 DEGs 取交集得到 30 个目标基因。通过 DAVID 数据库分析，作者发现目标基因富集于规律血管收缩、血管紧张素激活信号通路、3′-磷酸腺苷 5′-磷酸硫酸盐、轴突导引、伤口愈合、细胞对镉离子的反应、细胞锌离子稳态、细胞对铜离子的反应、细胞对锌离子的反应、生长负调节、铜离子解毒等 BP；富集于细胞连接、细胞皮层、细胞外液、胞外外体、细胞质、胞质溶胶等 CC；富集于锌离子结合 MF。KEGG 通路富集分析显示：目标基因参与血管平滑肌收缩和矿物吸收通路。作者选择了按 PPI 网络连接度排序的前 10 个 Hub 基因，接着进行 TCGA 数据库验证后，并对与 TCGA 数据库表达一致的 Hub 基因进行 ROC 曲线和 Kaplan-Meier 绘图仪的分析，鉴定出其中 4 个基因是结直肠癌中的枢纽基因，经过 ROC 曲线分析发现基因 MYL9、ACTG2 和 PDGFRA 对结直肠癌诊断均有一定准确性，基因 AGT 对结直肠癌诊断具有较高准确性。经过 K-M 曲线分析发现，PDGFRA、ACTG2 和 MYL9 与结直肠癌的预后显著相关。PDGFRA、ACTG2 和 MYL9 低表达均显示患者预后较好。在结直肠癌中，MYL9、ACTG2、AGT 和 PDGFRA 可能作为潜在的分子研究靶点。虽然本研究结果与以往的研究有所不同，但本研究为结直肠癌的诊断、治疗和预后提供了一些参考方向。而差异可能是由于分析的案例数量不同，不同的识别标准和实验错误，或其他原因。

PDGFRA 作为 PDGFA、PDGFB 和 PDGFC 的细胞表面受体，在调节胚胎发育、细胞增殖、存活和趋化性方面发挥重要作用。本研究分析发现，PDGFRA 在结直肠癌中低表达（$\log_2$FC＝－2.03），而且低表达与结直肠癌患者预后良好相关，有多项相关研究显示，PDGFRA 基因与结直肠癌的发生发展有密切关系[21-23]。这可能是因为 PDGFRA 是胃肠道黏膜正常发育所必需的酶，在骨髓间充质干细胞分化中起重要作用，而且其会根据环境，促进或抑制细胞增殖和细胞迁移，也是在胚胎发育过程中正常骨骼发育和头部关闭所要求的物质。

ACTG2 是高度保守的蛋白质，参与各种类型的细胞运动，并普遍表达在所有真核细胞。本研究分析发现，ACTG2 在结直肠癌中低表达（$\log_2$FC＝－3.03），而且其低表达与结直肠癌患者预后良好相关，这与赵林钢等[24]的研究相一致。有研究报道，此基因还与肝癌[25]、膀胱癌[26]和乳腺浸润性导管癌[27]的发生发展有密切关系。

MYL9 包含 EF-hand 结构域，是肌球蛋白调节亚基，通过其磷酸化在调节平滑肌和非肌细胞收缩活动中起重要作用，其涉及胞质分裂、受体封闭和细胞运动。本研究分析发现，MYL9 基因在结直肠癌中低表达（$\log_2$FC＝－3.99），而且低表达与结直肠癌患者预后良好相关。有研究报道，MYL9 基因在肺癌、乳腺癌、前列腺癌和恶性黑色素瘤等多种不同类型恶性肿瘤中均有异常表达[28-29]。尚少有 MYL9 基因与结直肠癌相关性的报道，此基因可供后续继续深入研究。

AGT 是内源性配体，是肾素-血管紧张素系统的基本成分，其是血压、体液和电解质稳态的有效调节剂。本研究发现，AGT 基因在结直肠癌中高表达（$\log_2$FC＝7.97）。有研究报道，AGT 基因与高尿酸血症[30]、糖尿病肾病[31]和妊娠高血压[32]相关。此基因的突变与对原发性高血压的易感性相关联，并能引起肾小管发育不全、肾小管发展的严重障碍。在 KEGG 的基因解析中显示 AGT 基因的缺陷与非家族的结构性心房纤颤和炎症性肠病相关。此前尚少有 AGT 基因与结直肠癌相关性的报道，此基因可供后续继续深入研究。

ACTG2 和 MYL9 基因高表达与结直肠癌患者预后不良相关，本研究发现，其在癌症通路中富集(hsa04270)，基因数量最多($P=0.004$)，ACTG2 和 MYL9 基因通过调控血管平滑肌收缩通路来影响结直肠癌的发生发展。血管平滑肌是一种高度特化的细胞，其主要功能是收缩。收缩时，血管平滑肌细胞缩短，从而减少血管直径，以调节血流量和压力。调节血管平滑肌细胞收缩状态的主要机制是胞浆钙离子(Ca^{2+})浓度([Ca^{2+}]c)的变化。为了响应血管收缩刺激，Ca^{2+} 从细胞内储存和(或)细胞外液动员起来，以增加血管平滑肌细胞中的[Ca^{2+}]c。[Ca^{2+}]c 的增加反过来激活 Ca^{2+}-CaM-MLCK 途径并刺激 MLC20 磷酸化，导致肌球蛋白-肌动蛋白相互作用，从而增强收缩力。收缩肌丝或 MLC20 磷酸化对 Ca^{2+} 的敏感性可能受其他信号通路的次级调节。在受体刺激过程中，肌球蛋白磷酸酶的抑制作用大大增强了收缩力。Rho/Rho 激酶、PKC 和花生四烯酸在这种增强中起着关键作用。介导松弛的信号事件包括去除收缩激动剂(被动松弛)和在持续存在收缩激动剂(主动松弛)的情况下激活环核苷酸依赖性信号通路。主动松弛是通过抑制血管平滑肌细胞中 Ca^{2+} 的动员和肌丝 Ca^{2+} 的敏感性而发生的。

本研究从 3 个 GEO 系列中鉴定出了结直肠癌和正常结直肠上皮细胞之间的 30 个 DEGs。根据 PPI 网络的连接度，确定了前 10 个 Hub 基因，然后通过 TCGA 数据库进行验证，接着对与 TCGA 数据库表达一致的 Hub 基因进行 ROC 曲线和 Kaplan-Meier 绘图仪的生存曲线分析，作者鉴定出其中 4 个基因是结直肠癌中的枢纽基因，这些基因包括 PDGFRA、ACTG2、MYL9 和 AGT，将为结直肠癌的研究提供一些新方向。

参考文献

[1] 黄波，李慧雯，常诚. miR-26b-5p 通过靶向 NFE2L3 调控结直肠癌细胞生长、迁移和侵袭[J]. 实用医学杂志，2022，38(2)：160-167.

[2] 姚菲，黄启友，黄晓颖，等. miR-33a-3p 通过调控 EphA2 影响结直肠癌化疗耐药[J]. 华中科技大学学报，2022，51(1)：7-13.

[3] SIEGEL R L，MILLER K D，JEMAL A. Cancer statistics，2020[J]. CA Cancer J Clin，2020，70(1)：7-30.

[4] 刘宗超，李哲轩，张阳，等. 2020 全球癌症统计报告解读[J/CD]. 肿瘤综合治疗电子杂志，2021，7(2)：1-13.

[5] 赵志娟，孟莲，刘春霞. 基于融合基因作用于横纹肌肉瘤的 miRNA-mRNA 调控网络的生物信息学分析[J]. 吉林大学学报(医学版)，2022，48(1)：154-162.

[6] 王晓萌，孟莲，李春森，等. 横纹肌肉瘤相关差异 miRNAs 和靶基因的生物信息学分析[J]. 重庆医科大学学报，2021，46(10)：1242-1247.

[7] CHEN J R，LIU C，CEN J M，et al. KEGG-expressed genes and pathways in triple negative breast cancer：protocol for a systematic review and data mining[J]. Medicine(Madr)，2020，99(18)：e19986.

[8] 冯勤超，邹贤，王国瑞，等. 基于微阵列数据分析的甲状腺癌 circRNA-miRNA 调控预测模型研究[J]. 南京医科大学学报(自然科学版)，2020，40(8)：1140-1148.

[9] LEWIS B P，BURGE C B，BARTEL D P. Conserved seed pairing，often flanked by adenosines，indicates that thousands of human genes are microRNA targets[J]. Cell，2005，120(1)：15-20.

[10] WONG N，WANG X W. miRDB：an online resource for microRNA target prediction and functional annotations[J]. Nucleic Acids Res，2015，43：D146-D152.

[11] AGARWAL V，BELL G W，NAM J W，et al. Predicting effective microRNA target sites in mammalian mRNAs[J]. Elife，2015，4：e05005.

[12] HUANG D W，SHERMAN B T，TAN Q N，et al. David bioinformatics resources：expanded annotation database and novel algorithms to better extract biology from large gene lists[J]. Nucleic Acids Res，2007，35：W169-W175.

[13] SZKLARCZYK D，GABLE A L，LYON D，et al. STRING v11：protein-protein association networks with increased coverage，supporting functional discovery in genome-wide experimental datasets[J]. Nucleic Acids Res，2019，47(D1)：D607-D613.

[14] 王子璐，刘立民，孙晓. 有氧运动对人骨骼肌基因表达的影响[J]. 中国医科大学学报，2020，49(6)：556-560.

[15] SHANNON P，MARKIEL A，OZIER O，et al. Cytoscape：a software environment for integrated models of biomolecular interaction networks[J]. Genome Res，2003，13(11)：2498-2504.

[16] 刘慧，陆嘉骏，樊璐，等. 基于网络药理学和生物信息学的辛伐他汀分子生物学机制研究[J]. 中国临床药理学杂志，2019，35(14)：1510-1513.

[17] GYÖRFFY B，LANCZKY A，EKLUND A C，et al. An online survival analysis tool to rapidly assess the effect of 22，277 genes on breast cancer prognosis using microarray data of

1,809 patients[J]. Breast Cancer Res Treat, 2010,123(3):725-731.
[18] 方宇,王海娟,李征洋,等. PD-L1 和 A2aR 在大肠癌中的表达及意义[J]. 河北医药,2021,43(3):345-348.
[19] 许涛,余阳,方军,等. CDX-2、NKD1、GDF11 在大肠癌中的表达及其与预后的相关性分析[J]. 湖南师范大学学报(医学版),2022,19(1):18-22.
[20] 许涛,余阳,方军,等. CDX-2、PD-L1、Ki-67 在大肠癌组织中的表达及与临床病理、预后的关系[J]. 中国临床研究,2022,35(5):639-643.
[21] 王梦莹,张军,吴星烨,等. 270 例胃肠间质瘤基因突变谱及与临床病理特征关系[J]. 重庆医科大学学报,2021,46(1):22-27.
[22] 徐国栋,张玉萍,张云香,等. 胸腺上皮性肿瘤中 PDGFRA 基因突变和蛋白表达的研究[J]. 诊断病理学杂志,2020,27(4):245-249.
[23] 吴小延,刘小云,古家美,等. 血小板源性生长因子受体 α 突变型胃肠间质瘤的临床病理特征及预后分析[J]. 中山大学学报(医学科学版),2021,42(4):603-612.
[24] 赵林钢,刘兆国,杨爱华. 水苏碱通过调控 ACTG2 蛋白表达抑制结肠癌生长[J]. 中国药学杂志,2018,53(13):1077-1082.
[25] 蒋晟昱,吴育,石美琴,等. 肌动蛋白 2 基因(ACTG2)对肝癌肿瘤细胞侵袭转移促进作用的实验研究[J]. 药学与临床研究,2019,27(1):10-15.
[26] 金敏,纪珊珊,徐珊珊,等. 基于共表达网络分析鉴定膀胱癌预后相关核心基因[J]. 华北理工大学学报,2022,24(1):36-42.
[27] 孙琰,李惠翔. 乳腺浸润性导管癌组织 ACTG2 和 MYH11 表达及预后关系[J]. 中华肿瘤防治杂志,2020,27(24):1977-1983.
[28] 覃聪昕,谭翔,王永勇,等. 沉默肌球蛋白轻链 9 基因对非小细胞肺癌 H1299 细胞珠增殖及迁移的影响[J]. 广西医学,2019,41(3):329-332.
[29] 游伊梦,刘庭波,沈建箴. 肌球蛋白轻链 9 在恶性肿瘤中的研究进展[J]. 中南大学学报(医学版),2021,46(10):1153-1158.
[30] 梁灼源,韦锋,欧阳楚君,等. 血管紧张素原基因多态性与高尿酸血症易感性的关系分析[J]. 中国医药科学,2021,11(23):195-198.
[31] 冉隆梅,王宇琴,赖红辉,等. ACE 及 AGT 基因多态性与糖尿病肾病发病风险的关系[J]. 中国实用医药,2019,14(27):196-197.
[32] 高英,王情,郭红,等. 外周血 ACE 基因 I/D、AGT 基因 M235T 多态性与妊娠期高血压疾病的关系[J]. 山东医药,2020,60(34):11-14.

(收稿日期:2023-11-07　修回日期:2024-02-16)

(上接第 1108 页)

Global Leadership Initiative on Malnutrition criteria on clinical outcomes of patients with gastric cancer[J]. JPEN J Parenter Enteral Nutr,2022,46(2):385-394.
[11] FUJIYA K,KAWAMURA T,OMAE K,et al. Impact of malnutrition after gastrectomy for gastric cancer on long-term survival[J]. Ann Surg Oncol,2018,25(4):974-983.
[12] LIANG Y,LIU L,XIE X,et al. Tumor size improves the accuracy of the prognostic prediction of lymph node-negative gastric cancer[J]. J Surg Res,2019,240:89-96.
[13] SLAGTER A E,TUDELA B,VAN AMELSFOORT R M,et al. Older versus younger adults with gastric cancer receiving perioperative treatment:results from the CRITICS trial[J]. Eur J Cancer,2020,130:146-154.
[14] YAMASHITA K,HOSODA K,NIIHARA M,et al. History and emerging trends in chemotherapy for gastric cancer[J]. Ann Gastroenterol Surg,2021,5(4):446-456.
[15] KARABULUT S,DOGAN I,USUL AFSAR C,et al. Does nutritional status affect treatment tolerability,chemotherapy response and survival in metastatic gastric cancer patients? Results of a prospective multicenter study in Turkey[J]. J Oncol Pharm Pract,2022,28(1):127-134.
[16] LIU C,LU Z,LI Z,et al. Influence of malnutrition according to the GLIM criteria on the clinical outcomes of hospitalized patients with cancer[J]. Front Nutr,2021,24(8):774636.
[17] ZHANG X,TANG T,PANG L,et al. Malnutrition and overall survival in older adults with cancer:a systematic review and meta-analysis[J]. J Geriatr Oncol,2019,10(6):874-883.
[18] FUJIYA K,KAWAMURA T,OMAE K,et al. Impact of malnutrition after gastrectomy for gastric cancer on long-term survival[J]. Ann Surg Oncol,2018,25(4):974-983.

(收稿日期:2023-10-27　修回日期:2023-11-26)