·基础研究·

肺腺癌关键基因筛选及CDC20在肺腺癌中的表达

周春雷[1,2],韩　菲[2],樊祥山[1*]

[1]南京大学医学院附属鼓楼医院病理科,江苏　南京　210008;[2]南京医科大学附属儿童医院病理科,江苏　南京　210008

[摘　要]　**目的:**筛选肺腺癌的关键基因,揭示肺腺癌潜在的分子机制。**方法:**利用3个肺腺癌相关基因芯片数据集,共包括156例肺腺癌和106例正常肺组织,进行肺腺癌潜在关键基因的筛选。**结果:**通过芯片数据分析得到341个重叠差异表达基因(differentially expressed gene,DEG),DEG主要富集的生物学过程是细胞外基质组织、调控化学激酶介导的信号通路、细胞对转化生长因子刺激的反应、细胞外基质组装、调控新生血管生成等。通过蛋白互作网络(protein protein interaction,PPI)筛选出10个关键基因,其中8个基因AURKA(aurora kinase A)、CDC20(cell division cycle 20)、CDH5(cadherin 5)、COL1α1(collagen Ⅰ,α1)、EDN1(endothelin 1)、MMP9(matrix metalloprotein 9)、PECAM1(platelet endothelial cell adhesion molecule 1)、SPP1(secreted phosphoprotein 1)与肺腺癌预后相关。其中CDC20与肺腺癌预后相关性最高,亚组分析发现,CDC20在肺腺癌各年龄段患者中的表达均明显高于正常组,CDC20在一期肺腺癌中表达低于二期、三期、四期,CDC20在腺泡型、实性为主型、混合型、透明细胞型表达高于其他亚型,CDC20在吸烟的肺腺癌患者中的表达低于不吸烟者。免疫组化检测显示CDC20在肺腺癌组织中表达明显高于正常肺组织,且腺泡型、实体型表达高于其他亚型。**结论:**CDC20在肺腺癌中高表达,且其高表达与肺腺癌的不良预后显著相关。CDC20有可能成为肺腺癌治疗的新靶点。

[关键词]　肺腺癌;CDC20;Gene Expression Omnibus(GEO);关键基因;预后

[中图分类号]　R734.2　　[文献标志码]　A　　[文章编号]　1007-4368(2020)01-032-07

doi:10.7655/NYDXBNS20200107

Screening of key genes by bioinformatics analysis and the expression of CDC20 in lung adenocarcinoma

ZHOU Chunlei[1,2], HAN Fei[2], FAN Xiangshan[1*]

[1]Department of Pathology, Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School, Nanjing 210008; [2]Department of Pathology, the Affiliated Children's Hospital of Nanjing Medical University, Nanjing 210008, China

[**Abstract**]　**Objective**: The aim of this study was to identify the key genes and uncover the potential molecular mechanisms of lung adenocarcinoma(LAC). **Methods**: In our study, three microarray data sets of LAC genes, including 156 cases of LAC and 106 cases of normal lung tissue, were used to screen the potential key genes. **Results**: First, 341 overlapping differentially expressed genes(DEG) were found by microarray data analysis. The main biological processes of DEG enrichment were extracellular matrix tissue, regulation of chemical kinase-mediated signaling pathways, cell response to transformed growth factor stimulation, extracellular matrix assembly, and regulation of angiogenesis. Next, 10 key genes were screened out by using protein protein interaction(PPI) network, including 8 genes AURKA(aurora kinase A), CDC20(cell division cycle 20), CDH5(cadherin 5), COL1α1(collagen I, α1), EDN1(endothelin 1), MMP9(matrix metalloprotein 9), PECAM1(platelet endothelial cell adhesion molecule 1), SPP1(secreted phosphoprotein 1), which were correlated with the prognosis of LAC. Additional, CDC20 has the highest correlationship with the prognosis of LAC. The subgroup analysis showed that expression of CDC20 in all age groups of LAC patients increased significantly compared with nomal control; expression of CDC20 in tumor stage Ⅰ is lower than tumor stage Ⅱ, Ⅲ, Ⅳ; expression of CDC20 in aciner subtype, solid subtype, mixed subtype, clear cell subtype was higher than other subtypes; expression of CDC20 in LAC patients with smoking was lower than the non-smokers. Furthermore, the results by immunohistochemical showed that expression of CDC20 in LAC tissues was

[基金项目]　南京医科大学科技发展基金(NMUB2018086)

*通信作者(Corresponding author),E-mail:fxs23@163.com

significantly higher than that in normal lung tissues, in acinar subtype and solid subtype was higher than other subtypes. **Conclusion**: CDC20 is high expressed in LAC, which is correlated with poor prognosis of LAC. CDC20 might be a potential biomarker and therapeutic target for LAC.

[**Key words**] lung adenocarcinoma; CDC20; Gene Expression Omnibus(GEO); key genes; prognosis

[J Nanjing Med Univ, 2020, 40(01): 032-038]

肺癌是目前全球死亡率最高的恶性肿瘤之一,肺腺癌是肺癌中最常见的病理学亚型,几乎占全部肺癌的40%~50%[1]。近年来在肺癌治疗方面取得了进展,但肺腺癌的预后仍较差,5年以上生存率不足18%[2]。肺腺癌的发生发展是一个多因素的过程,有多种基因参与这个过程[3-4]。因此,肺腺癌关键基因的鉴定对了解肺腺癌的发病机制具有重要意义,并能够为治疗提供可能的新靶点。

蛋白质组学、基因组学和生物信息学,特别是基因芯片的快速发展,为我们探索肿瘤的发病机制和关键基因提供了便利。美国国立生物信息技术中心(NCBI)的基因芯片公共数据库GEO(http://www.ncbi.nih.gov/geo/)是当今最大、最全面且公开的基因表达谱数据库,可以存档和自由分发由科学界提交的全套微阵列、新一代测序和其他形式的高通量功能基因组数据[5]。本研究从GEO数据库分析3个肺腺癌相关基因芯片数据集GSE32863、GSE74706、GSE43458,使用在线分析软件GEOR2分析差异表达基因(differentially expressed gene, DEG),对在3个数据集中都存在的差异基因进行功能富集分析、KEGG通路分析,构建蛋白互作网络(PPI),利用Cytoscape软件分析网络中的关键基因,并使用Kaplan-Meier在线工具(http://kmplot.com/analysis/),进一步分析这些关键基因与肺腺癌预后的相关性,并对与肺腺癌预后相关性最高的基因进行功能分析[6],以期为肺腺癌的早期诊断和治疗找到更准确、可靠的新靶点。

1 材料和方法

1.1 材料

从GEO上检索得到肺腺癌与正常肺组织或邻近非肿瘤肺组织的基因表达谱,分别是GSE43458、GSE32863、GSE74706。GSE43458的芯片数据基于GPL6244,包括80例肺腺癌和30例正常肺组织。GSE32863数据集基于GPL6884平台,包括58例肺腺癌和58例相邻的正常肺组织。GSE74706数据集基于GPL13497平台,包括18例肺腺癌和18例正常肺组织。数据分析使用GEO的在线工具GEOR2,利用韦恩图分析在3个数据集中都有的DEG(DEG的标准是$P<0.05$,|lgFC|>1,FC即变化倍数fold change)。

1.2 方法

1.2.1 DEG的筛选和分析

使用Enrichr对DEG进行功能注释、功能富集分析和KEGG通路分析,使用STRING(http://string-db.org)[7]对DEG构建PPI网络。利用Cytoscape软件平台[8]重现PPI网络,并且使用cytoHubba筛选网络中degree最高的10个基因。

1.2.2 关键基因与肺腺癌预后的关系

利用在线工具Kaplan-Meier Plot(http://kmplot.com/analysis/)进一步分析10个基因是否与肺腺癌预后相关,挑选出与肺腺癌预后最相关的基因。利用UALCAN分析TCGA数据库该基因在不同性别、年龄以及其他分类标准的肺腺癌患者中的表达情况。使用LinkedOmics在线数据库分析与该基因表达正相关或负相关的基因,并且对相关基因的功能及通路进行富集分析。

1.2.3 免疫组化检测CDC20在肺腺癌中表达水平

收集2018年4月—2019年4月本院肺腺癌患者的石蜡组织标本,以癌旁正常组织为对照。采用免疫组化法检测样本中CDC20(CDC20抗体,杭州联科生物技术股份有限公司)蛋白的表达,对免疫组化的结果判读及打分采用国际通用H-score法,将阳性程度和阳性范围综合考虑打分:细胞显色0分(阴性)、1分(弱阳性)、2分(阳性)、3分(强阳性),计算公式为H-score=1×弱染色细胞百分率+2×中度染色细胞百分率+3×强染细胞百分率(10个高倍镜视野)。观察癌组织和正常组织CDC20阳性表达情况,分析CDC20表达情况和肺腺癌患者临床病理特征之间的关系。

1.3 统计学方法

采用Graphpad Prism 5和SPSS 21.0软件分析处理实验数据。定量数据采用均数±标准差($\bar{x}\pm s$)表

示，两组间比较采用t检验，多组间比较采用单因素方差分析，两两比较采用LSD-t检验。生存分析采用Kaplan-Meier和Log-rank检验法。$P \leqslant 0.05$为差异有统计学意义。

2 结 果

2.1 肺腺癌中DEG的鉴定

我们从GSE32863、GSE74706、GSE43458这3个数据集中分别得到1 239、4 189、897个DEG（$P < 0.05$，$|lgFC| > 1$），并得到341个重叠DEG（图1）。DEG的功能富集情况和KEGG通路分析见图2。

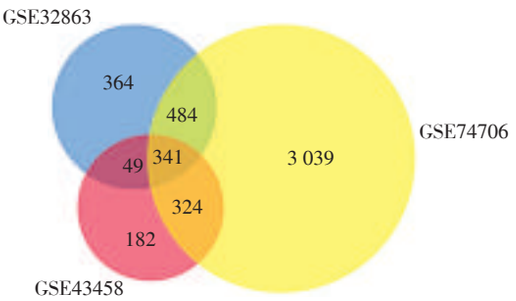


图1 GSE32863、GSE74706、GSE43458数据集中的肺腺癌DEG

Figure 1 Identification of overlapping DEGs in GSE32863, GSE74706 and GSE43458 of lung adenocarcinoma

2.2 PPI的构建及关键候选基因的筛选

通过STRING和Cytoscape软件将所有重叠的341个DEG构建到PPI网络中（图3）。我们使用cytoHubba筛选出网络中degree得分最高的10个基因MMP9、VWF、PECAM1、CD34、CDH5、COL1A1、SPP1、EDN1、AURKA、CDC20。利用UALCAN分析数据库中这些基因在肺腺癌中的表达情况，结果显示与正常肺组织相比，10个关键基因在肺腺癌中都存在差异表达。

2.3 关键基因与肺腺癌预后的相关性

使用Kaplan-Meier在线工具进一步验证关键基因与肺腺癌的相关性，结果显示8个基因AURKA、CDC20、CDH5、COL1A1、EDN1、MMP9、PECAM1、SPP1与肺腺癌预后相关。其中CDC20[HR=2.39（1.87~3.05），log-rank $P=8.6\times10^{-13}$]相关性最高（图4）。

2.4 CDC20在不同性别、年龄以及其他分类标准的肺腺癌患者中的表达情况

利用UALCAN分析TCGA数据库中CDC20基因在不同性别、年龄以及其他分类标准的肺腺癌患者中的表达情况，结果显示CDC20在肺腺癌患者各年龄段的表达均明显高于正常组，差异有统计学意义（$P < 0.001$），且随着年龄的增加表达下降；CDC20在不同人种及性别的肺腺癌患者中的表达均明显高于正常组，差异有统计学意义（$P < 0.01$，$P < 0.001$），但各人种及性别间的表达无明显差异；CDC20在肺腺癌患者不同分期的表达均明显高于正常组，差异有统计学意义（$P < 0.001$），且在一期肺腺癌患者中表达低于二期、三期、四期患者；CDC20在各肺腺癌亚型中的表达均明显高于正常组，差异有统计学意义（$P < 0.05$，$P < 0.001$），其中在混合型、透明细胞型、实性为主型、腺泡型中表达高于其他亚型；CDC20在吸烟的肺腺癌患者中的表达低于不

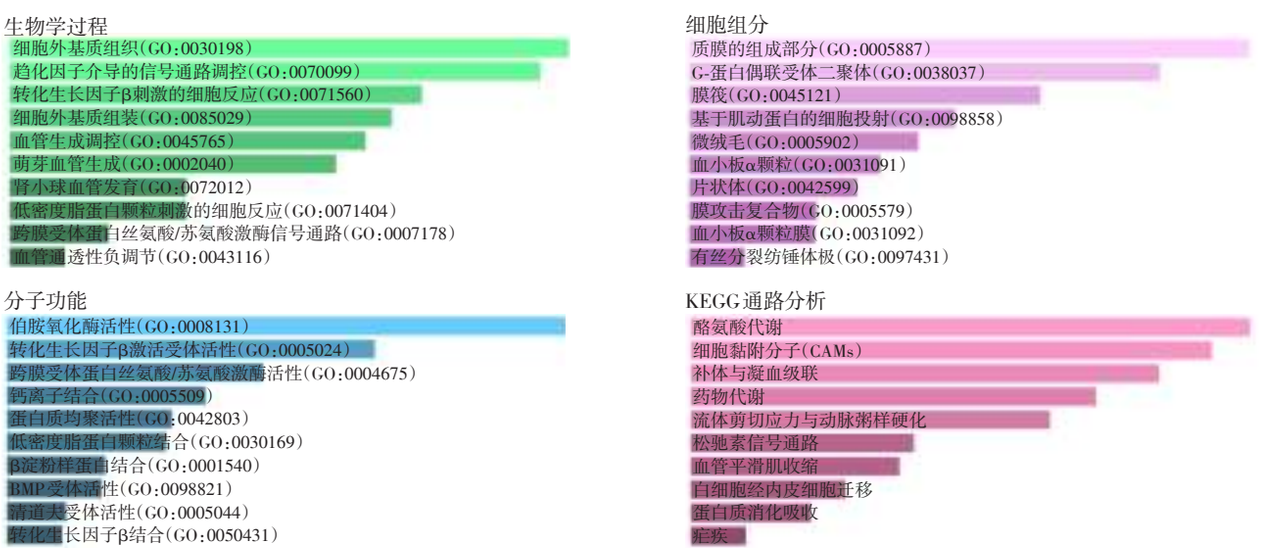


图2 DEG的功能富集和KEGG通路分析

Figure 2 Gene ontology analysis and KEGG-pathway analysis of DEG

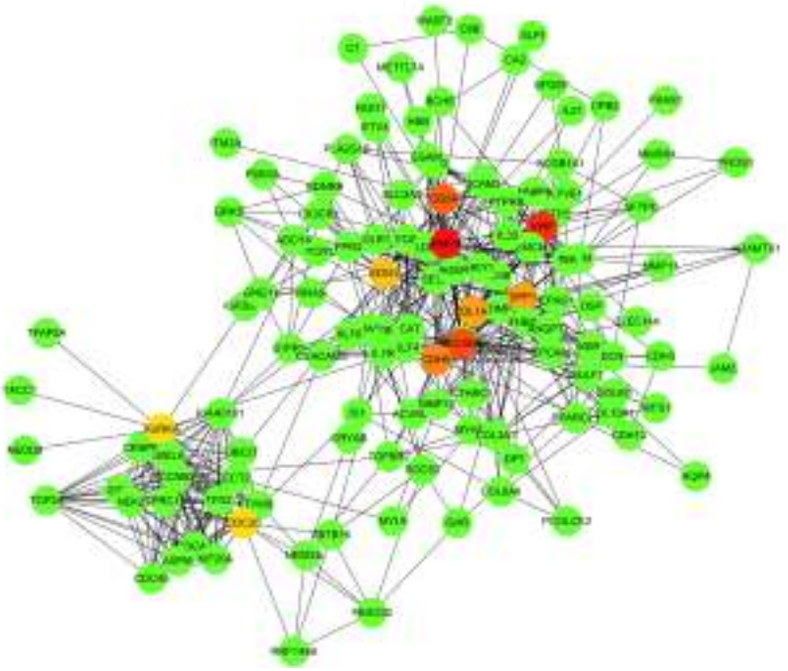


图3　PPI网络及筛选的关键基因

Figure 3　The hub genes identified from the PPI network

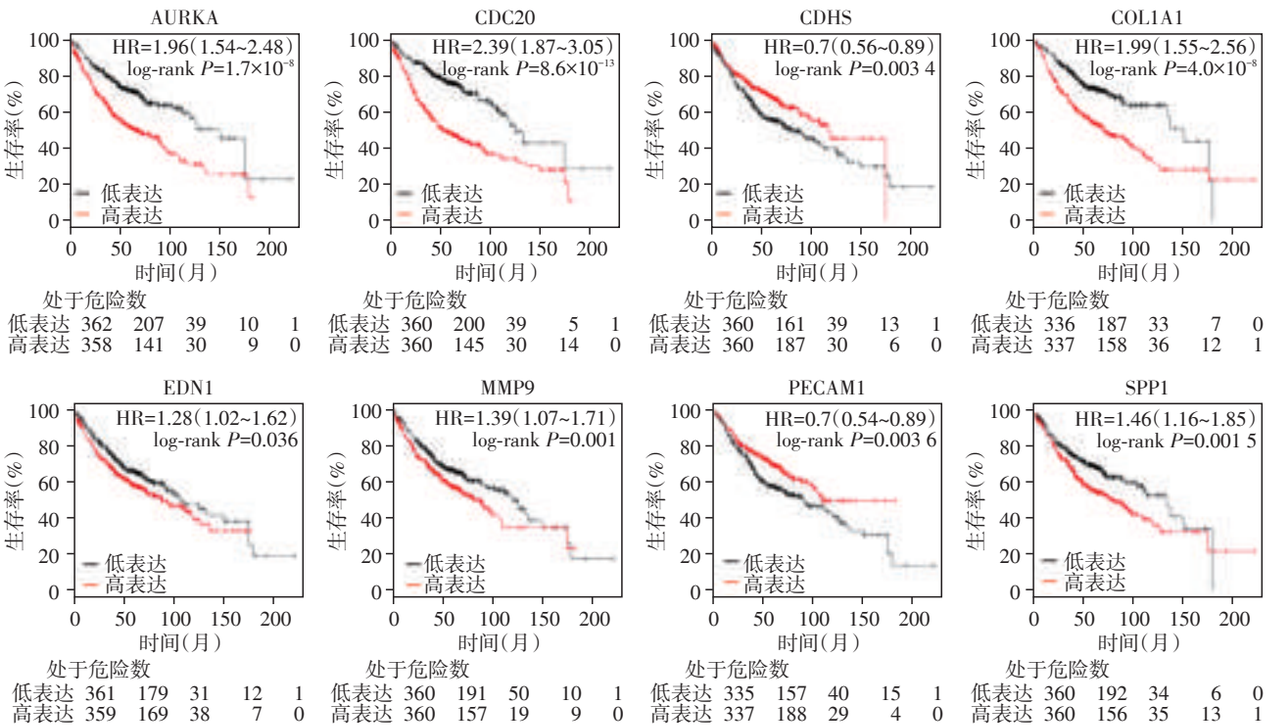


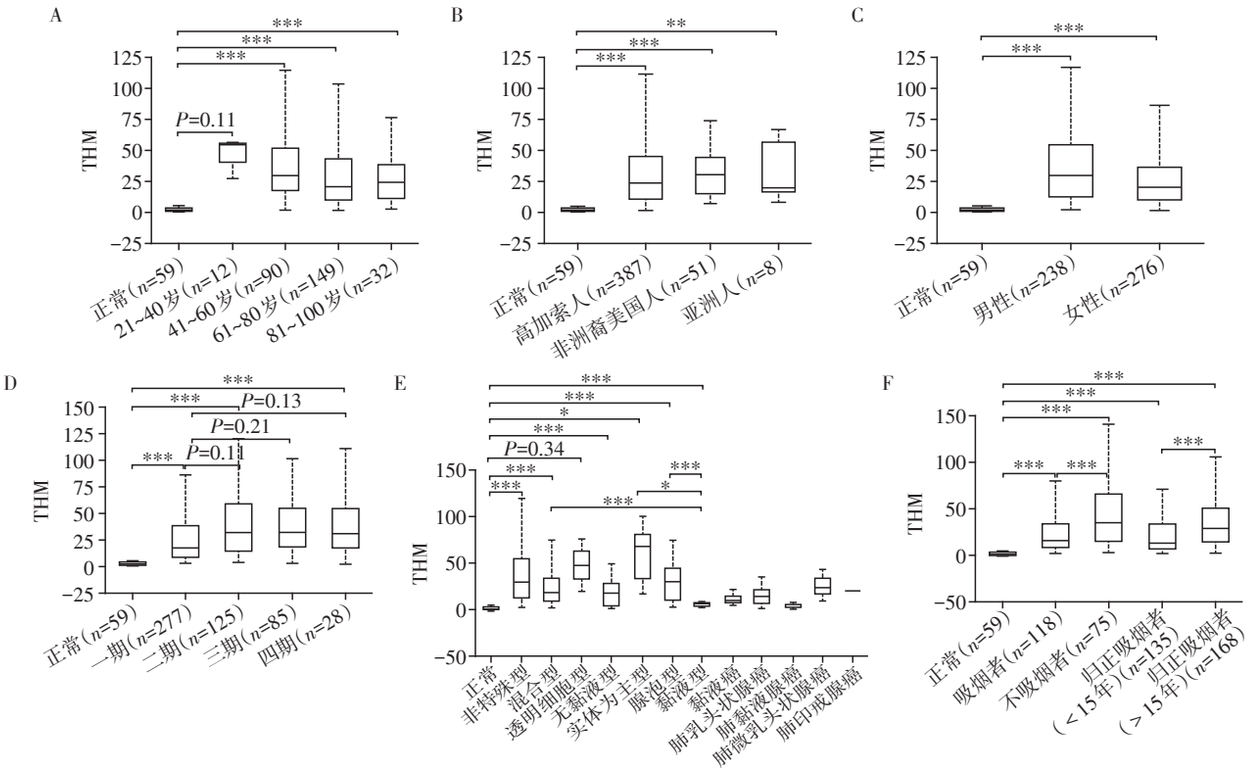
图4　与肺腺癌预后相关的关键基因与生存关系

Figure 4　The prognosis value of hub genes in the Kaplan-Meier Plotter analysis

吸烟者,差异有统计学意义($P<0.001$,图5)。

2.5　免疫组化检测CDC20在肺腺癌中的表达水平

为了验证CDC20在肺腺癌中的表达情况,及其与病理学特征之间的关系,我们通过免疫组化方法检测了肺腺癌与正常肺组织CDC20的表达情况。结果显示CDC20在肺腺癌组织中表达明显高于正常肺组织,差异有统计学意义($P<0.001$);进一步,我们分析了不同亚型肺腺癌CDC20的表达差异,结果显示在肺腺癌中,腺泡型、实体型CDC20的表达明显高于其他亚型,差异有统计学意义($P<0.01$,图6)。



A：不同年龄肺腺癌患者CDC20的表达；B：不同种族肺腺癌患者CDC20的表达；C：不同性别肺腺癌患者CDC20的表达；D：不同分期肺腺癌患者CDC20的表达；E：不同病理亚型肺腺癌患者CDC20的表达；F：不同吸烟习惯肺腺癌患者CDC20的表达。THM：transcript per million；两组比较，$^{*}P<0.05$，$^{**}P<0.01$，$^{***}P<0.001$。

图5 CDC20在不同分类肺腺癌患者中的表达情况

Figure 5 The expression of CDC20 in different subgroups of lung adenocarcinoma patients

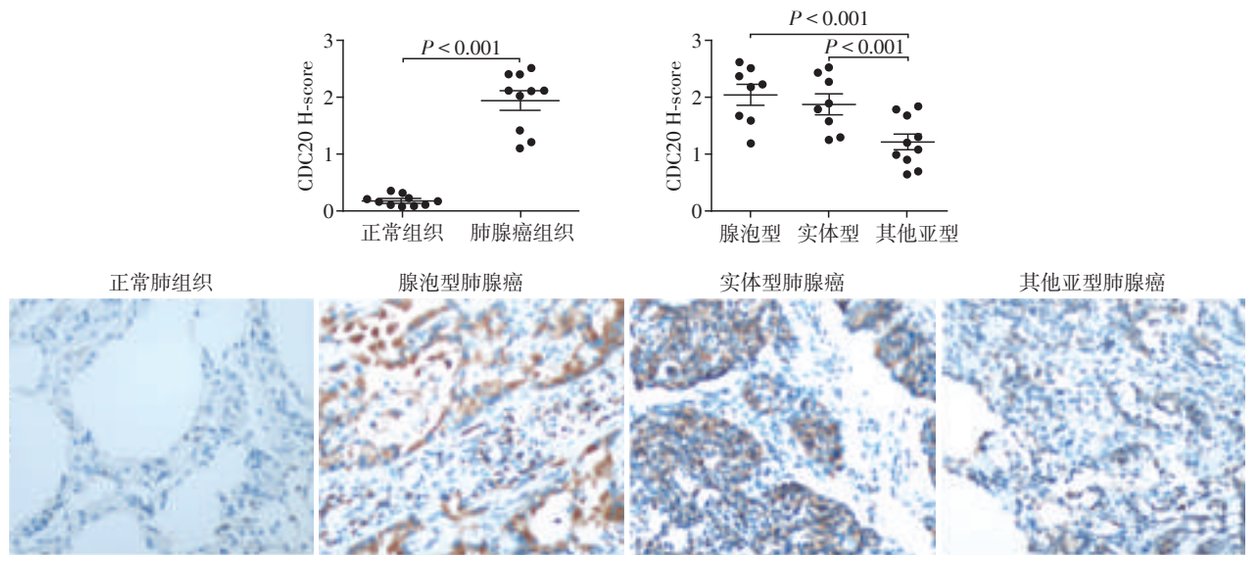


图6 免疫组化检测CDC20在肺腺癌中表达水平(×200)

Figure 6 The expression of CDC20 in lung adenocarcinoma by IHC(×200)

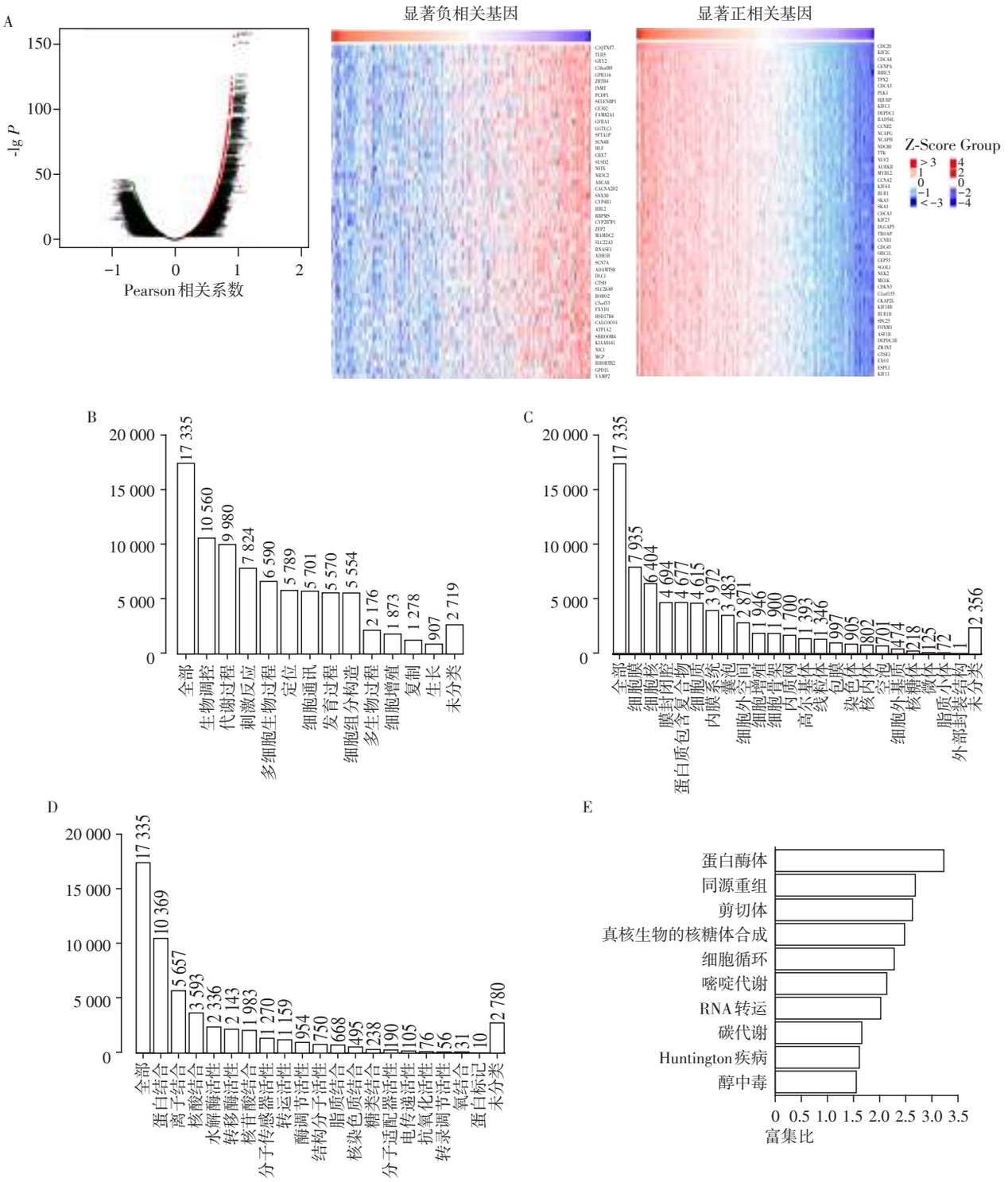
2.6 与CDC20表达正相关或负相关的基因及其功能分析

使用LinkedOmics在线数据库分析与CDC20基因表达正相关或负相关的基因，其中CDC20表达相关基因主要富集的生物学过程是生物调控、代谢过程、刺激反应、多细胞生物过程、定位、细胞通讯等；富集的细胞组分是细胞膜、细胞核、膜封闭腔、蛋白质包含复合物、细胞质等；分子功能主要富集在蛋

白结合、离子结合、核酸结合、水解酶激活、转移酶激活等；相关的KEGG通路富集在蛋白酶体、同源重组、剪接体、细胞周期、嘧啶代谢、RNA运输(图7)。

3　讨　论

近年来，为探索肺腺癌发生发展的内在机制，



A：CDC20相关基因；B~E：CDC20相关基因的功能分析：生物学过程的富集(B)、细胞组分的富集(C)、分子功能的富集(D)、KEGG信号通路分析(E)。

图7　CDC20表达相关基因及其功能分析

Figure 7　The associated genes of CDC20 in lung adenocarcinoma and Gene Ontology analysis of CDC20

开展了大量的基础研究和临床研究，对肺腺癌的治疗也有了较大的进步，但全球肺腺癌的发病率和死亡率仍然居高不下[9]。造成这一困境的主要原因是肺腺癌的发生发展是一个多因素的过程，受众多基因的影响，但大多数研究只关注单个队列人群或单个基因。这可能会限制结果的准确性和可信度。本研究结合3个GEO基因芯片数据，进行多数据库多角度分析验证，增加了研究结果的准确性和可信度。本研究最终发现CDC20是肺腺癌中的关键基因，且在各种亚型的肺腺癌中的表达均明显升高。

首先，通过对3个GEO基因芯片数据进行分析，得到341个重叠DEG。在由341个DEG构建的PPI网络中，我们使用cytoHubba 筛选出网络得分最高的10个基因使用Kaplan-Meier在线工具进一步验证关键基因与肺腺癌的相关性，结果显示8个基因AURKA、CDC20、CDH5、COL1A1、EDN1、MMP9、PECAM1、SPP1与肺腺癌预后相关，其中CDC20相关性最高。因此我们将CDC20作为最关键的核心基因做进一步分析。

CDC20是细胞周期相关蛋白之一，最初发现时被认为是调控细胞分裂活动的关键蛋白[10]。进一步深入研究显示，CDC20在肿瘤的发生发展过程中具有重要作用。CDC20的异常表达可导致有丝分裂发生错误，并引起一些癌基因过表达和抑癌基因的突变或者丢失从而能够促进肿瘤细胞的增殖，抑制肿瘤细胞的凋亡[11-12]。为了深入研究CDC20在肺腺癌中的作用，我们分析CDC20在不同性别、年龄以及其他分类标准的肺腺癌患者中的表达情况，发现CDC20高表达与肺腺癌显著相关，且与肺腺癌的临床病理特征密切相关。为了验证以上结果，我们检测了肺腺癌患者的组织中CDC20的表达，结果显示与前期的生物信息学结果相一致。进一步我们对CDC20相关基因进行分析及功能进行富集，分析发现相关基因主要富集的生物学过程是生物调控、代谢过程、刺激反应、多细胞生物过程等，相关的KEGG通路富集在蛋白酶体、同源重组、剪接体、细胞周期等。上述结果提示CDC20参与多种生物学过程，其在肺腺癌的发生发展中具有重要作用，有可能成为肺腺癌治疗的新靶点。

[参考文献]

[1] SIEGEL RL, MILLER KD, JEMAL A. Cancer statistics, 2019[J]. CA Cancer J Clin, 2019, 69(1): 7-34

[2] EBERLE A, JANSEN L, CASTRO F, et al. Lung cancer survival in Germany: A population-based analysis of 132, 612 lung cancer patients[J]. Lung Cancer, 2015, 90(3): 528-533

[3] DEVARAKONDA S, MORGENSZTERN D, GOVINDAN R. Genomic alterations in lung adenocarcinoma[J]. Lancet Oncol, 2015, 16(7): E342-E351

[4] 罗 娟，周 晓，程志祥. FoxO1在非小细胞肺癌中的研究进展[J]. 南京医科大学学报(自然科学版)，2018，38(8)：1161-1166

[5] TANG YC, ZHANG ZX, TANG YC, et al. Identification of potential target genes in pancreatic ductal adenocarcinoma by bioinformatics analysis[J]. Oncol Lett, 2018, 16(2): 2453-2461

[6] GYORFFY B, SUROWIAK P, BUDCZIES J, et al. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer[J]. PLoS One, 2013, 8(12): e82241

[7] SZKLARCZYK D, FRANCESCHINI A, WYDER S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life[J]. Nucleic Acids Res, 2015, 43(Database issue): D447-D452

[8] SMOOT ME, ONO K, RUSCHEINSKI J, et al. Cytoscape 2.8: new features for data integration and network visualization[J]. Bioinformatics, 2011, 27(3): 431-432

[9] CALVAYRAC O, PRADINES A, PONS E, et al. Molecular biomarkers for lung adenocarcinoma[J]. Eur Respir J, 2017, 49(4): 1601734

[10] YUAN I, LEONTIOU I, AMIN P, et al. Generation of a spindle checkpoint arrest from synthetic signaling assemblies[J]. Curr Biol, 2017, 27(1): 137-143

[11] KIM Y, CHOI JW, LEE JH, Et al. MAD2 and CDC20 are upregulated in high-grade squamous intraepithelial lesions and squamous cell carcinomas of the uterine cervix [J]. Int J Gynecol Pathol, 2014, 33(5): 517-523

[12] WANG LX, ZHANG JF, WAN LX, et al. Targeting Cdc20 as a novel cancer therapeutic strategy [J]. Pharmacol Ther, 2015, 151: 141-151

[收稿日期] 2019-04-11