论著·基础研究

食管鳞状细胞癌基因组芯片生物信息学分析及靶向药物预测

李　倩，高境泽，李　云，宋　堃，沈倩诚
上海交通大学基础医学院医药生物信息学中心，上海 200025

[摘要] **目的**·探究食管鳞状细胞癌（esophageal squamous cell carcinoma，ESCC）的发生机制及其潜在靶向药物，为诊断和治疗ESCC提供理论依据。**方法**·选取2个GEO集（GSE38129、GSE20347），用R语言筛选差异表达基因（differentially expressed genes，DEGs）并进行GO（Gene Ontology）和KEGG（Kyoto Encyclopedia of Genes and Genomes）富集分析。对DEGs行蛋白质相互作用（protein-protein interaction，PPI）网络分析，获得最显著模块基因以及关键基因，并对关键基因磷酸化酶B激酶（phosphorylase B kinase，*PBK*）做靶向药物预测。**结果**·2个数据集共包含670条相同的DEGs，其中下调基因342条、上调基因328条。GO和KEGG富集分析结果显示，DEGs主要富集到细胞外结构的组织、细胞外基质的组织、p53信号通路、IL-17信号通路、细胞周期等通路。通过对PPI网络做密集度分析，共筛选出20条关键基因。其中，关键基因*PBK*与细胞周期相关，在2个数据集中表达量均有上调；通过变构位点探测和化合物库虚拟筛选，预测出了*PBK*的潜在药物Compound 1。**结论**·通过生物信息学的方法能够有效分析ESCC的关键基因。关键基因*PBK*的靶向药物预测结果可能为ESCC的靶向治疗提供一定的参考。

[关键词] 食管鳞状细胞癌；差异表达基因；GO富集分析；KEGG富集分析；蛋白质相互作用网络；靶向药物预测
[DOI] 10.3969/j.issn.1674-8115.2020.02.008　[中图分类号] R735.1　[文献标志码] A

Bioinformatics analysis of esophageal squamous cell carcinoma genomic chip and prediction of targeted drug

LI Qian, GAO Jing-ze, LI Yun, SONG Kun, SHEN Qian-cheng
Medicinal Bioinformatics Center, Shanghai Jiao Tong University College of Basic Medical Sciences, Shanghai 200025, China

[Abstract] Objective·To explore the mechanism of esophageal squamous cell carcinoma (ESCC) and its potential targeted drugs, and to provide the theoretical basis for diagnosis and treatment of ESCC. **Methods**·Two GEO sets GSE38129 and GSE20347 were selected, and differentially expressed genes (DEGs) were screened by R language. GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment analysis were conducted for DEGs. The most significant module genes and key genes were analyzed by protein-protein interaction (PPI) network for DEGs. Targeted drug prediction was made for phosphorylase B kinase (*PBK*). **Results**·A total of 670 DEGs were identified, consisting of 342 down-regulated genes and 328 up-regulated genes. The enriched functions and pathways of DEGs included extracellular structure organization, extracellular matrix organization, p53 signaling pathway, IL-17 signaling pathway and cell cycle. Twenty key genes were identified by analyzing DEGs' PPI network. The key gene *PBK* was related to the cell cycle, and the expression of *PBK* was up-regulated in the two data sets. The potential drug Compound 1 of *PBK* was predicted by allosteric site detection and compound library virtual screening. **Conclusion**·The key genes of ESCC can be effectively analyzed by bioinformatics. The prediction results of targeted drugs of key gene *PBK* may provide reference for the targeted therapy of ESCC.

[Key words] esophageal squamous cell carcinoma (ESCC); differentially expressed gene (DEG); GO enrichment analysis; KEGG enrichment analysis; protein-protein interaction (PPI) network; targeted drug prediction

食管鳞状细胞癌（esophageal squamous cell carcinoma，ESCC）是一种常见的恶性肿瘤，具有较高的发病率和死亡率。临床上，该疾病仅在出现某些症状后才能被确诊，且预后较差。因此，理解ESCC的发生机制、探寻其早期检测的生物标志物并开展靶向药物预测对于该疾病的诊断及治疗十分重要。目前，传统药物因存在选择性差、毒副作用强、易产生耐药性等问题，使得其临床应用效果不佳。近年来，变构药物因毒性弱、选择性好等特点引起了越来越多研究者的关注，但针对其变构位点的研发仅借助实验手段则未能获得较好的结果。因此，本研究拟通过生物信息学的方法对变构位点进行预测。然而，以往针对ESCC的生物信息学分析存在数据集单一、数据样本量少

[作者简介] 李　倩（1989—），女，实验技术员，硕士；电子信箱：liq297@163.com。
[通信作者] 沈倩诚，电子信箱：youarefree.1986@163.com。
[Corresponding Author] SHEN Qian-cheng, E-mail: youarefree.1986@163.com.

等问题，且很少就其分析结果开展靶向药物预测等更深入的研究。基于此，本研究以基因表达综合数据库（Gene Expression Omnibus，GEO，https://www.ncbi.nlm.nih.gov/geo）下载的数据集为材料进行生物信息学分析，筛选出与 ESCC 发生密切相关的关键基因，并对该关键基因做进一步的靶向药物预测，从而识别其潜在的变构位点，为 ESCC 的靶向药物研发提供一定的参考。

1 材料与方法

1.1 数据获取

GEO 是存储高通量基因表达数据、芯片和微阵列的一个公共数据库，隶属于美国国立生物技术信息中心（National Center for Biotechnology Information，NCBI）[2]。本研究从 GEO 中下载获得 2 个数据集 GSE38129、GSE20347，其均来自人类 ESCC 组织与正常组织的 mRNA 阵列。GSE38129 共包含 60 组样本，30 组为正常组织，其余 30 组为 ESCC 组织。GSE20347 共包含 34 组样本，17 组为正常组织，其余 17 组为 ESCC 组织。

1.2 差异表达基因的筛选

使用 R 语言 limma 包筛选 2 个数据集中正常组织和 ESCC 组织的差异表达基因（differentially expressed genes，DEGs）。基因表达的差异用 P 值和差异倍数（fold change，FC）的对数（logFC）表示。$P<0.05$ 表示差异具有统计学意义。本研究将 $P<0.05$ 且 $|logFC|>1$ 的基因视为 DEGs。

1.3 蛋白质相互作用网络的构建及关键基因的筛选

使用在线数据库 STRING（https://string-db.org）对组织中蛋白质间的相互作用进行分析，构建 DEGs 的蛋白质相互作用（protein-protein interaction，PPI）网络。采用 Cytoscape 软件对 PPI 网络进行可视化分析，并使用 Cytoscape 的 MCODE 插件对 PPI 网络进行密集度分析，筛选出最显著的模块[3-4]。随后，使用 Cytoscape 的 CytoHubba 插件，用最大团中心性（maximal clique centrality，MCC）方法根据打分值的高低筛选出排名前 20 的关键基因[5]，用于后续开展进一步的靶向药物预测。

1.4 GO 和 KEGG 富集分析

用 R 语言 clusterProfiler 包对 DEGs 做 GO（Gene Ontology）和 KEGG（Kyoto Encyclopedia of Genes and Genomes）功能富集[6]，分析其涉及的相关通路，富集分析的结果以参数 $P<0.05$ 作为入选标准。

1.5 靶向药物预测

AlloSitePro（http://mdl.shsmu.edu.cn/AST/）是一种基于口袋特征和微扰模型来预测蛋白变构位点的在线网站，亦是一种便携的变构工具，可为不同蛋白质及感兴趣的复合物中的各种变构效应研究提供帮助[7]。本研究使用 AlloSitePro 预测蛋白的潜在变构位点，而后使用 Schrodinger 软件对筛选得到的变构位点进行小分子虚拟筛选，以获得能够结合在变构位点上的小分子化合物，实现对基于结构的变构药物设计的靶向预测。

2 结果

2.1 DEGs 分析

本研究运用 R 语言对数据集 GSE38129、GSE20347 的 DEGs 进行筛选，结果显示，前者共筛选出 785 条 DEGs，后者共筛选出 1 061 条 DEGs；2 个数据集共有 670 条相同的 DEGs，其中 342 条为下调基因、328 条为上调基因（图 1）。

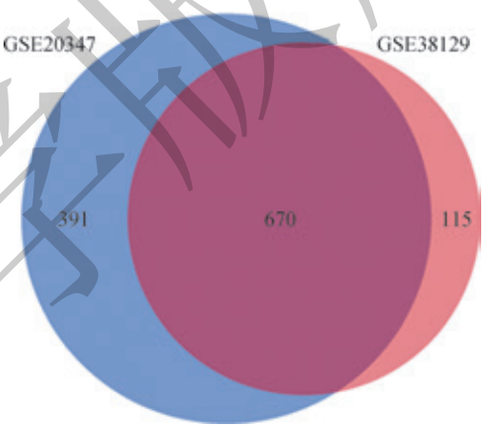
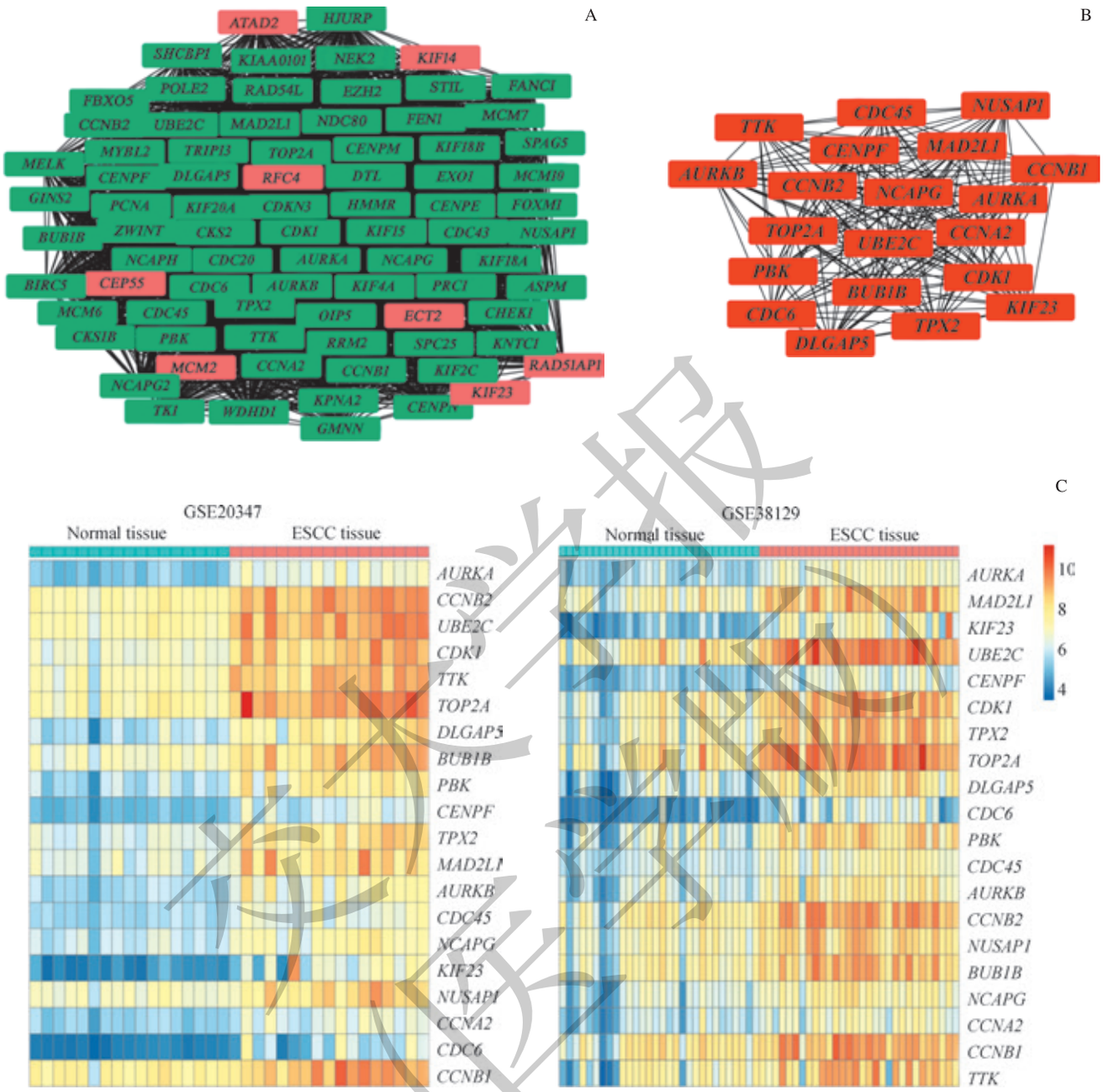


图 1 2 个数据集的 DEGs 的 Venn 图
Fig 1 Venn diagram of DEGs in the two datasets

2.2 PPI 网络构建及关键基因筛选

本研究通过将 DEGs 输入数据库 STRING，构建 PPI 网络；并用 Cytoscape 的 MCODE 插件对 PPI 网络进行分组，形成多个模块，最终筛选出最显著模块（即评分最高模块）基因。随后，运用 Cytoscape 的 CytoHubba 插件对 PPI 网络进行分析，即使用 MCC 方法筛选出排名前 20 的 DEGs，记为关键基因；通过观察关键基因在 2 个数据集的热图发现，其在癌症组织的表达量均有所上调（图 2）。



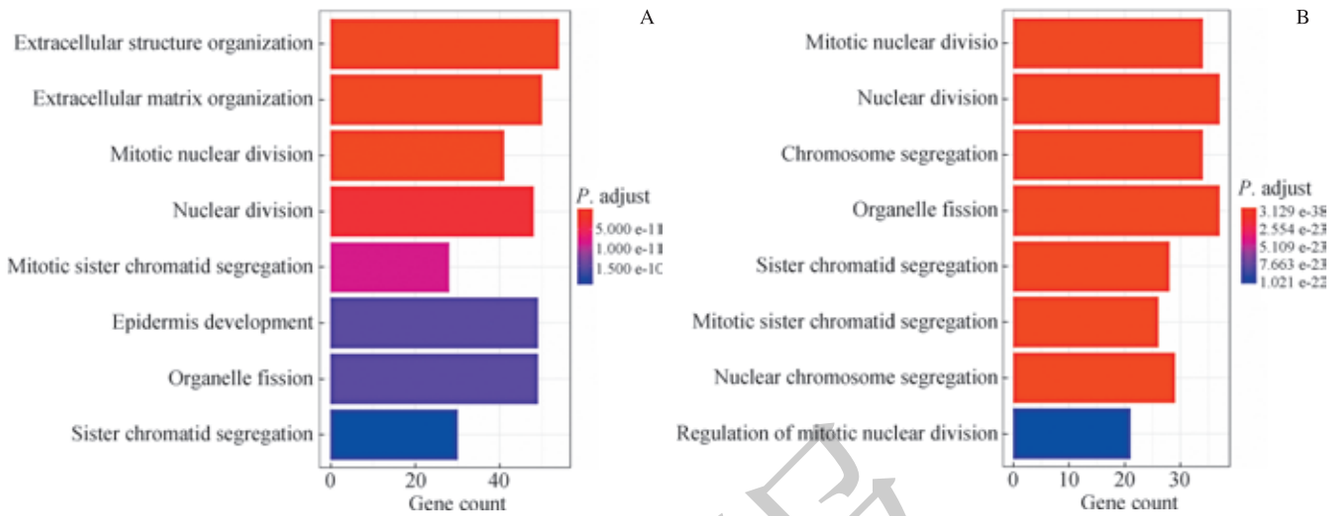
Note: A. PPI network of the most prominent module genes. Green—down-regulated genes, red—up-regulated genes, line—the connections between proteins. B. PPI network of top 20 key genes. Line—the connections between proteins. *TTK*—Thr/Tyr kinase, *CDC45*—cell division cycle 45, *NUSAP1*—nucleolar and spindle associated protein 1, *CENPF*—centromere protein F, *MAD2L1*—mitotic arrest deficient 2 like 1, *AURKB*—aurora kinase B, *CCNB2*—cyclin B2, *NCAPG*—non-SMC condensin Ⅰ complex subunit G, *AURKA*—aurora kinase A, *CCNB1*—cyclin B1, *TOP2A*—DNA topoisomerase Ⅱ alpha, *UBE2C*—ubiquitin-conjugating enzyme E2 C, *CCNA2*—cyclin A2, *PBK*—phosphorylase B kinase, *BUB1B*—BUB1 mitotic checkpoint serine/threonine kinase B, *CDK1*—cyclin-dependent kinases 1, *CDC6*—cell division cycle 6, *DLGAP5*—DLG associated protein 5, *TPX2*—TPX2 microtubule nucleation factor, *KIF23*—kinesin family member 23. C. Heatmaps of key genes in the two datasets. Red—up-regulation of gene expression level, green—down-regulation of gene expression level, shades of color—the degree of change in the expression.

图 2 最显著模块基因的 PPI 网络和关键基因的 PPI 网络及热图分析

Fig 2 Analysis of the PPI network of the most prominent module genes and the PPI network and the heatmaps of key genes

2.3 GO 和 KEGG 富集分析

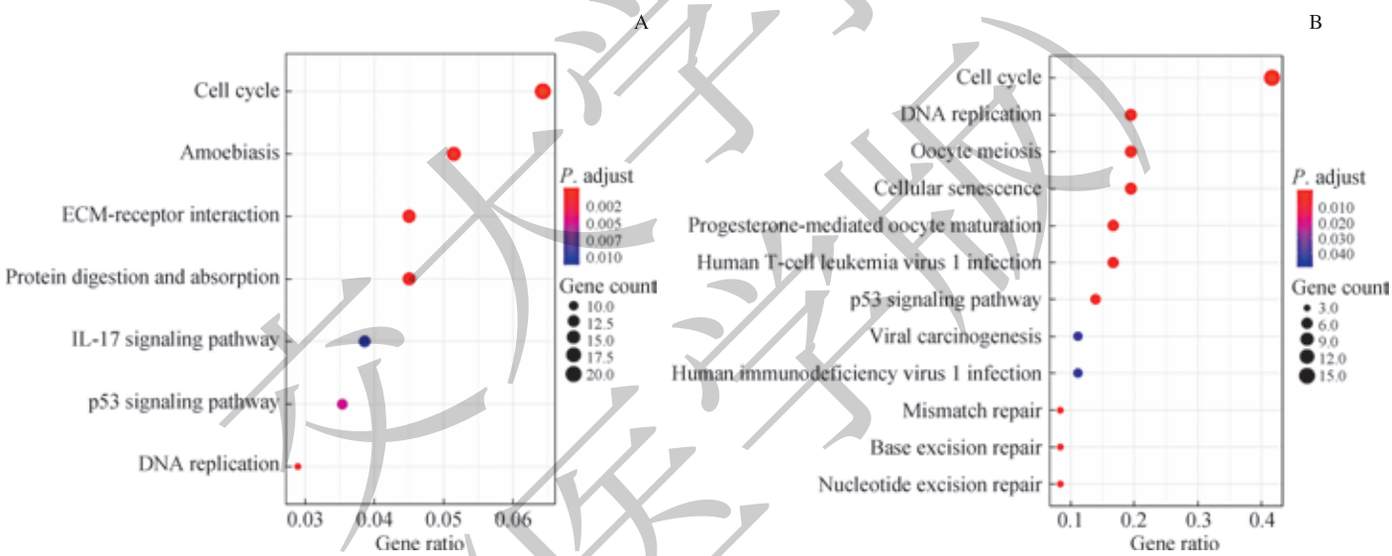
本研究使用 GO 富集分析，以 $P<0.05$ 作为阈值，发现 DEGs 富集于细胞外结构的组织、细胞外基质的组织、有丝核分裂、核分裂等通路；最显著模块基因的 GO 富集分析主要富集到有丝核分裂、细胞器裂变、染色体隔离等通路（图 3）。同时，我们使用 KEGG 对 DEGs 进行富集分析，结果发现主要富集于细胞周期、ECM- 受体相互作用、p53 信号通路、IL-17 信号通路、DNA 复制等通路；最显著模块基因的 KEGG 富集分析主要富集到以下通路，包括细胞周期、p53 信号通路、DNA 复制（图 4）。



Note: A. GO enrichment analysis of DEGs. B. GO enrichment analysis of the most prominent module genes in PPI network.

图 3 DEGs 及 PPI 网络最显著模块基因的 GO 富集分析

Fig 3 GO enrichment analysis of DEGs and the most prominent module genes in PPI network



Note: A. KEGG enrichment analysis of DEGs. B. KEGG enrichment analysis of the most prominent module genes in PPI network. The bubble size represented the number of genes enriched. Gene ratio referred to the ratio of DEGs related to this pathway.

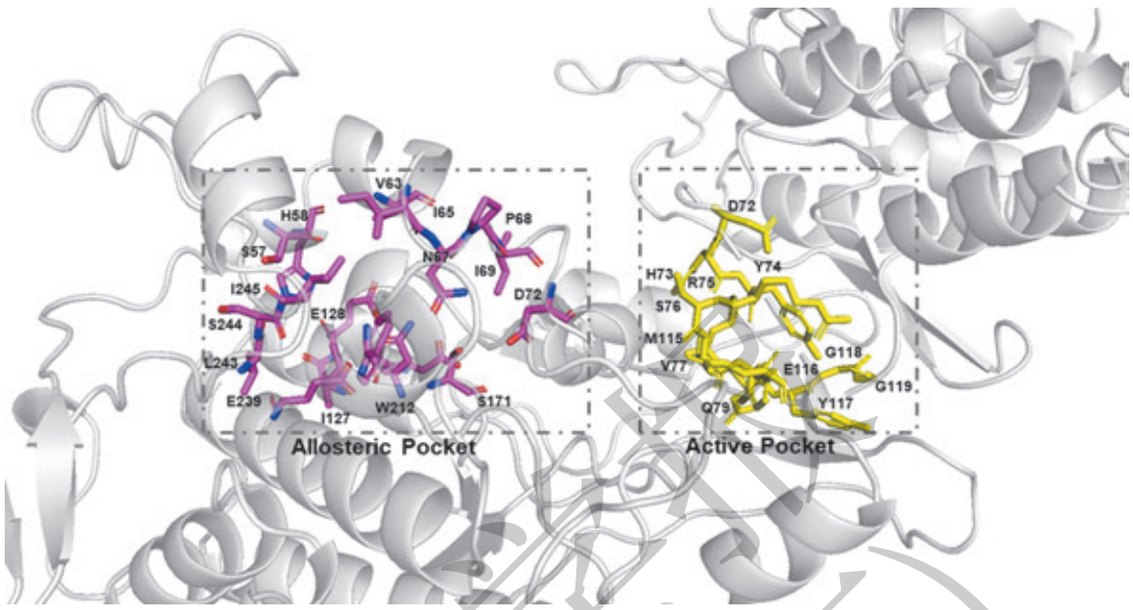
图 4 DEGs 及 PPI 网络最显著模块基因的 KEGG 富集分析

Fig 4 KEGG enrichment analysis of DEGs and the most prominent module genes in PPI network

GO 和 KEGG 富集分析的结果显示，*PBK* 是与细胞周期通路相关的基因。在 ESCC 中过表达的 *PBK* 可能促进肿瘤细胞增殖，导致 ESCC 患者生存率下降，被认为是 ESCC 潜在的治疗靶点[8]；同时，*PBK* 在其他癌症如肺癌、乳腺癌、膀胱癌等多种类型癌症中的表达均有上调[9-12]。*PBK* 既属于最显著模块中的基因，又属于关键基因，GO 和 KEGG 富集分析的结果显示 *PBK* 被富集到了细胞周期等与癌症相关的通路，且通过观察热图发现 *PBK* 在 2 个数据集中表达量均上调；继而推断，*PBK* 在 ESCC 的发生、发展中发挥着重要的作用。

2.4 靶向变构药物预测

本文选取 AlloSitePro 预测结果中打分最高的变构位点进行基于分子对接的虚拟筛选，图 5 显示了 *PBK* 的活性位点及预测的潜在变构位点。

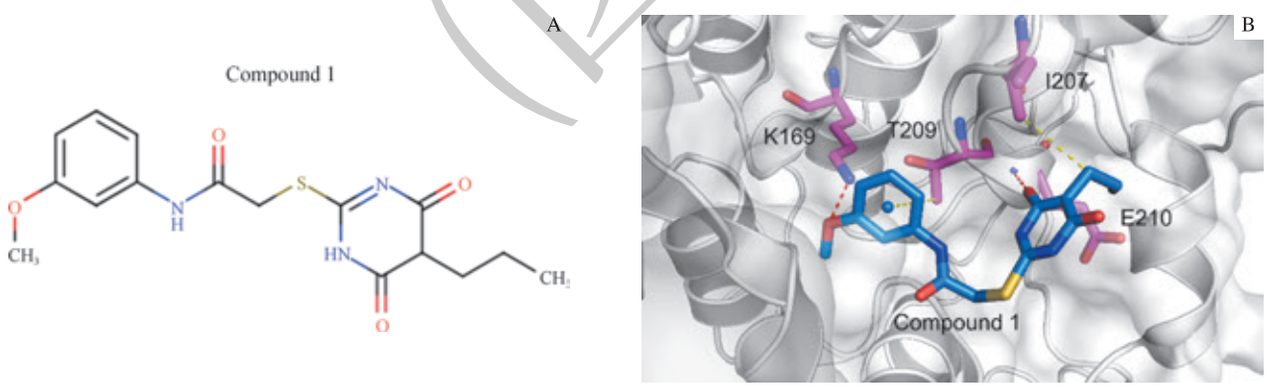


Note: Allosteric site and active site of *PBK* were displayed as sticks and highlight in magenta and yellow, respectively. Allosteric site was surrounded by several residues such as Ser57—S57, His58—H58, Val63—V63, Ile65—I65, Asn67—N67, Pro68—P68, Ile69—I69, Asp72—D72, Ile127—I127, Glu128—E128, Ser171—S171, Trp212—W212, Glu239—E239, Leu243—L243, Ser244—S244 and Ile245—I245. Active site was consisted of several residues such as Asp72—D72, His73—H73, Tyr74—Y74, Arg75—R75, Ser76—S76, Val77—V77, Gln79—Q79, Met115—M115, Glu116—E116, Tyr117—Y117, Gly118—G118 and Gly119—G11.

图 5 *PBK* 的活性位点及由 AlloSitePro 预测的潜在变构位点

Fig 5 Active site of *PBK* and potential allosteric site predicted by AlloSitePro

本研究用 Schrodinger 软件的 Glide 模块获取了打分最高的 100 个化合物，其中化合物 1（Compound 1）打分为 −7.05 分，结构如图 6A 所示。我们利用 PLIP（protein-ligand interaction profiler）软件[13]分析化合物的结合模式发现，该化合物甲氧基上的氧可以与残基 K169 的侧链形成氢键，嘧啶二酮上的羰基可以与残基 E210 的主链形成氢键；同时，该化合物的苯环和正丙基可以与残基 T209、I207 的侧链形成疏水作用（图 6B）。因此，通过上述相互作用的研究表明，Compound 1 可能与 *PBK* 的潜在变构位点靶向结合，是一种潜在的靶向 *PBK* 的变构药物。



Note: A. Structure of Compound 1. B. Targeted binding of Compound 1 with *PKB* potential allosteric site residues. Lys169—K169, Thr209—T209, Ile207—I207, Glu20—E210.

图 6 潜在变构化合物及其与 *PBK* 的疏水作用

Fig 6 Potential allosteric compound and its hydrophobic interaction with *PBK*

3 讨论

在过去的几十年里，微阵列技术和生物信息学分析被广泛应用于基因突变筛查、肿瘤发生的相关基因及通路研究以及治疗靶点的筛选等。本研究通过对 GEO 数据集 GSE38129、GSE20347 进行分析，筛选出 670 条 DEGs；其中包含基质金属蛋白酶 3（matrix metalloproteinase 3，*MMP3*）、*MMP9*、*MMP13*、*MYBL2*（MYB proto-oncogene like 2）、*COL11A1*（collagen type Ⅺ alpha 1 chain）、*CHEK1*（checkpoint kinase 1）等，在 ESCC 发生与发展中扮演着重要角色。研究 [14] 显示，*MMP3*、*MMP9* 和 *MMP13* 在肿瘤的侵袭转移中起着关键作用。*MYBL2* 是 ESCC 的一个重要致癌基因，可以促进细胞的增殖和转移 [15]。*COL11A1* 可通过 ECM- 受体相互作用通路参与 ESCC 的发生与发展，可作为治疗 ESCC 的靶基因。*CHEK1*（又名 *CHK1*）是细胞周期的关键检查点 [16]，在卵巢癌、肺癌等多种癌症中过度表达，被认为是癌症治疗的潜在目标 [17-18]。

对 DEGs 的 GO 和 KEGG 富集分析结果显示，前者主要富集到细胞外结构的组织、细胞外基质的组织、有丝核分裂、核分裂等通路，后者则主要富集到细胞周期、ECM- 受体相互作用、p53 信号通路、IL-17 信号通路、DNA 复制等通路。研究 [19] 显示，IL-17A(IL-17) 信号通路可促进肿瘤的进展。细胞周期是细胞进行分裂和复制的过程，与细胞的增殖密切相关，不受控制的细胞增殖是癌症的特征之一 [20]。对 DEGs 构建 PPI 网络，用 Cytoscape 的 MCODE 插件对 PPI 网络进行分组，筛选出最显著模块基因，再用 Cytoscape 的 CytoHubba 插件筛选出 20 条关键基因。其中，*PBK* 既属于最显著模块基因又属于关键基因。通过对 GO 和 KEGG 富集分析做进一步分析显示，*PBK* 与细胞周期等与癌症相关的通路有关，且 *PBK* 在 2 个数据集中的表达量均上调。随后，本研究采用 AlloSitePro 算法对关键基因 *PBK* 的蛋白表面潜在变构位点进行预测并就化合物进行虚拟筛选，结果显示获得了潜在靶向 *PBK* 的变构分子 Compoud 1，为 ESCC 的靶向治疗提供了新的思路。

综上所述，本研究采用生物信息学的分析方法对在 ESCC 组织和正常组织的基因表达谱进行筛选，发现了可能参与 ESCC 发生与发展的 DEGs；随后，对该基因进行功能富集分析和蛋白互助网络分析，揭示出一些可能参与 ESCC 发病机制的富集通路和关键基因；并通过对关键基因的潜在靶向药物进行预测，实现对开发治疗 ESCC 药物的进一步探索。由于本研究仅进行了生物信息学分析而并未对分析结果开展试验验证，因此在未来的工作中可能需要通过扩大样本数量、开展进一步的试验来验证我们的推测。通过对 GEO 数据集的生物信息学分析我们发现，本研究结果或将为 ESCC 肿瘤发生机制的探索提供一定的帮助，关键基因的发现亦可能作为潜在的生物标志物用于临床 ESCC 的诊断与治疗。

参·考·文·献

[1] Song K, Li Q, Gao W, et al. AlloDriver: a method for the identification and analysis of cancer driver targets[J]. Nucleic Acids Res, 2019, 47(W1): W315-W321.

[2] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository[J]. Nucleic Acids Res, 2002, 30(1): 207-210.

[3] 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.

[4] Bandettini WP, Kellman P, Mancini C, et al. MultiContrast Delayed Enhancement (MCODE) improves detection of subendocardial myocardial infarction by late gadolinium enhancement cardiovascular magnetic resonance: a clinical validation study[J]. J Cardiovasc Magn Reson, 2012, 14(1): 83.

[5] Chin CH, Chen SH, Wu HH, et al. CytoHubba: identifying hub objects and sub-networks from complex interactome[J]. BMC Syst Biol, 2014, 8(Suppl 4): S11.

[6] Yu GC, Wang LG, Han YY, et al. ClusterProfiler: an R package for comparing biological themes among gene clusters[J]. OMICS, 2012, 16(5): 284-287.

[7] Song K, Liu X, Huang W, et al. Improved method for the identification and validation of allosteric sites[J]. J Chem Inf Model, 2017, 57(9): 2358-2363.

[8] Ohashi T, Komatsu S, Ichikawa D, et al. Overexpression of PBK/TOPK contributes to tumor development and poor outcome of esophageal squamous cell carcinoma[J]. Anticancer Res, 2016, 36(12): 6457-6466.

[9] Park JH, Lin ML, Nishidate T, et al. PDZ-binding kinase/T-LAK cell-originated protein kinase, a putative cancer/testis antigen with an oncogenic activity in breast cancer[J]. Cancer Res, 2006, 66(18): 9186-9195.

[10] Shih MC, Chen JY, Wu YC, et al. TOPK/PBK promotes cell migration *via* modulation of the PI3K/PTEN/AKT pathway and is associated with poor prognosis in lung cancer[J]. Oncogene, 2012, 31(19): 2389-2400.

[11] O Leary PC, Penny SA, Dolan RT, et al. Systematic antibody generation and validation *via* tissue microarray technology leading to identification of a novel protein prognostic panel in breast cancer[J]. BMC Cancer, 2013, 13: 175.

[12] Singh PK, Srivastava AK, Dalela D, et al. Expression of PDZ-binding kinase/T-LAK cell-originated protein kinase (PBK/TOPK) in human urinary bladder transitional cell carcinoma[J]. Immunobiology, 2014, 219(6): 469-474.

[13] Salentin S, Schreiber S, Haupt VJ, et al. PLIP: fully automated protein-ligand interaction profiler[J]. Nucleic Acids Res, 2015, 43(W1): W443-W447.

[14] Li YY, Zhou CX, Gao Y. Podoplanin promotes the invasion of oral squamous cell carcinoma in coordination with MT1-MMP and Rho GTPases[J]. Am J Cancer Res, 2015, 5(2): 514-529.

[15] Qin HD, Liao XY, Chen YB, et al. Genomic characterization of esophageal squamous cell carcinoma reveals critical genes underlying tumorigenesis and poor prognosis[J]. Am J Hum Genet, 2016, 98(4): 709-727.

[16] Zhang YW, Hunter T. Roles of Chk1 in cell biology and cancer therapy[J]. Int J Cancer, 2014, 134(5): 1013-1023.

[17] Paez-Pereda M, Kuchenbauer F, Arzt E, et al. Regulation of pituitary hormones and cell proliferation by components of the extracellular matrix[J]. Braz J Med Biol Res, 2005, 38(10): 1487-1494.

[18] Liu B, Qu JL, Xu FX, et al. MiR-195 suppresses non-small cell lung cancer by targeting *CHEK1*[J]. Oncotarget, 2015, 6(11): 9445-9456.

[19] Wang L, Yi TS, Kortylewski M, et al. IL-17 can promote tumor growth through an IL-6-Stat3 signaling pathway[J]. J Exp Med, 2009, 206(7): 1457-1464.

[20] Alabsi AM, Lim KL, Paterson IC, et al. Cell cycle arrest and apoptosis induction *via* modulation of mitochondrial integrity by Bcl-2 family members and caspase dependence in *Dracaena cinnabari*-treated H400 human oral squamous cell carcinoma[J]. Biomed Res Int, 2016, 2016: 4904016.

[收稿日期] 2019-06-17　　[本文编辑] 邢宇洋