

基于 TCGA 和 GEO 数据库——联合挖掘肾透明细胞癌 诊断和预后的生物标志物

高风霞¹ 王 珍² 高慧洁¹ (通讯作者)

1.河北工程大学附属医院 检验科, 河北 邯郸 056002

2.河北工程大学, 河北 邯郸 056002

摘要:目的 基于基因表达综合数据库(GEO)和肿瘤基因图谱(TCGA)数据库生物学信息,筛选肾透明细胞癌(ccRCC)中差异表达的关键基因,并分析其预后价值。方法 从 GEO 数据库下载 GSE40435 和 GSE66272 表达数据作为测试数据集,鉴定出差异常表达基因(DEGs)。利用生物本体论(GO)和京都基因与基因组百科全书(KEGG)富集分析,探讨 DEGs 在 ccRCC 中的生物学功能。使用 STRING 数据库构建 PPI 网络,利用 cytoHubba 识别出 Hub 基因。从 TCGA 数据库获取 ccRCC 表达数据作为验证数据集,初步验证 Hub 基因表达与 ccRCC 临床分期的相关性,在 ccRCC 组织和正常组织中的表达情况。最后,生成 KM 生存曲线和 ROC 曲线以验证 Hub 基因表达。结果 GSE40435 和 GSE66272 测试数据集共筛选出差异常基因 915 个,上调基因 396 个,下调基因 529 个。通过 PPI 网络构建和 Hub 基因分析,我们选择了前 10 个基因。利用 TCGA 数据集证实 10 个 Hub 基因,发现 6 个关键基因即 CEP55、ASPM、PTTG1、NUSAP1、KIF20A 和 PRC1,与 ccRCC 的早期诊断、肿瘤分期和不良预后有关。结论 ccRCC 中六个异常表达的关键基因对早期诊断、肿瘤分期确定和不良预后预测具有重要信息价值。

关键词: 肾透明细胞癌; GEO 数据库; TCGA 数据库; 基因本体论; 京都基因与基因组百科全书; 预后

肾透明细胞癌(clear cell renal cell carcinoma, ccRCC)是来源于肾实质小管上皮细胞的肿瘤,是肾癌的常见类型,约占全部肾癌的 75%^[1],肾透明细胞癌起病隐匿,不易发现。约有 15%的患者在诊断时已出现转移^[2]。当下,肾部分切除术或者根治性肾切除术能够对局部的这种癌症进行治疗。可是,在发生转移的情况下,其临床预后会比较差,并且化疗在转移性 ccRCC 的治疗方面效果不理想。因而,从基因水平上研究 ccRCC 疾病得发生、发展,寻找新的标志物,提高 ccRCC 疾病的预防和治疗至关重要。

1 材料与方法

1.1 数据来源

从基因表达综合数据库(gene expression omnibus, GEO)获取 ccRCC mRNA

(GSE40435 和 GSE66272)作为测试数据集,其中 GSE40435 包含 101 个癌组织样本和 101 个癌旁组织样本, GSE66272 包含 27 个癌组织样本和 27 个正常肾组织样本。肿瘤基因图谱(the cancer genome atlas, TCGA)数据库作为验证数据集。数据中有 532 例患者的癌组织数据以及 72 例癌旁组织数据。

1.2 方法

(1)差异表达基因(DEGs)的筛选

利用 Rv4.0.3 软件的 Limma 包(版本: 3.40.2)来研究 GSE40435 和 GSE66272 数据集的差异表达基因(differentially expressed genes, DEGs),将“调整后的 P 值<0.05 且 log₂(倍数变化)>1 或 log₂(倍数变化)<-1”为条件筛选 DEGs。将获得的 DEGs 使用 Biovenn 进行维恩图可视化分析。通过 Heatmapper 数据库挖掘和分析生成热图。

(2)DEGs 的 GO 和 KEGG 通路富集分析。

借助 Metascape 网站,针对筛选所得的 ccRCC 中的 DEGs,进行基因本体(gene ontology, GO)以及京都基因与基因组百科全书(Kyoto encyclopedia of genes and genomes, KEGG)通路的富集分析。

(3)预测 PPI 网络的构建

用 STRING 数据库构建蛋白质-蛋白质相互作用(protein-protein interaction, PPI)网络。利用 Cytoscape 3.10.3 软件对 PPI 网络做进一步的可视化处理,在 cytohubba 插件中进行模块化分析,进而找出显著的基因。

(4) Hub 基因表达验证

从 TCGA 数据库中选取 ccRCC 数据集, 利用 ggplot2 包分析 Hub 基因表达与临床分期之间的关系。Hub 基因表达水平在 ccRCC 癌组织与癌旁组织中的差异表达情况。

(5) Hub 基因对 ccRCC 患者生存的影响

使用 R 软件包 survival 和 survminer 进行 log rank 检验并绘制 Kaplan-Meier (KM) 生存曲线, 用于表示 Hub 基因在 ccRCC 患者之间的总生存 (OS) 分布。用 TCGA 数据库数据验证 Hub 基因表达和 ccRCC 预后之间可能存在的关系, 绘制受试者工作特征 (ROC) 曲线, 计算曲线下面积 (area under curve, AUC) 来评判其诊断能力。

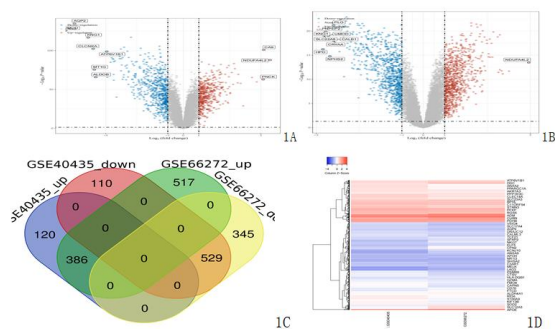
1.3 统计学处理

利用 R 软件 (4.0.3 版本) 进行统计学分析。用 t 检验分析比较 ccRCC 组织与癌旁组织间 Hub 基因表达水平的差异。采用 Survival 软件包 (3.2-10 版本) 对患者生存资料进行分析, 组间比较分析采用 Kaplan-Meier 法并行 Log-rank 检验; $P < 0.05$ 则差异有统计学意义。

2 结果

2.1 2 个微阵列数据集 DEGs 分析

通过分析 GSE40435 和 GSE66272 数据集, 鉴定出 GSE40435 506 个上调基因和 639 个下调基因, GSE66272 中鉴定出 903 个上调基因和 874 个下调基因, 见图 1A 和 1B。并将两个数据集获得的 DEGs 进行维恩图可视化分析, 显示共有 915 个重叠的 DEGs, 上调基因 396 个, 下调基因 529 个, 见图 1C。将 DEGs 用 Heatmapper 生成热图, 见图 1D。



(1A, 1B) 两个火山图显示 GSE40435 和 GSE66272 中所有表达的基因。(1C) Venn 图展示重叠的 DEGs。(1D) 热图显示了 915 个重叠的 DEGs。蓝色表示下调基因, 红色表示上调基因。

图 1 ccRCC 差异表达基因的筛选。

2.2 DEGs 的 GO 和 KEGG 通路富集分析。

DEGs 进行 GO 和 KEGG 富集分析显示, DEGs 在多个生物通路中显著富集。BP 术语 (图 2A), DEGs 主要富集于羧酸代谢过程; 小分子分解代谢过程; CC 术语 (图 2B), DEGs 主要富集于细胞的顶端部分; 基底质膜; MF 术

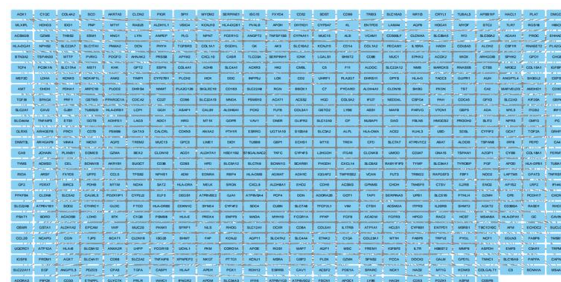
语的富集分析 (图 2C) 显示, DEGs 富集于氧化还原酶活性; 维生素结合; KEGG 通路 (图 2D) 显示, DEGs 富集于碳代谢; 吞噬小体。



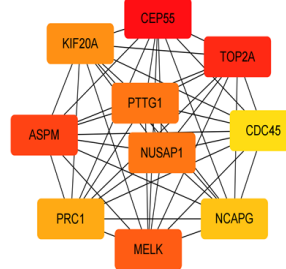
BP(图 2A); CC(图 2B); MF (图 2C); KEGG (图 2D)

图 2 差异表达基因的 KEGG 和 GO 富集分析

2.3 Hub 基因的 PPI 网络构建与分析。



3A



3B

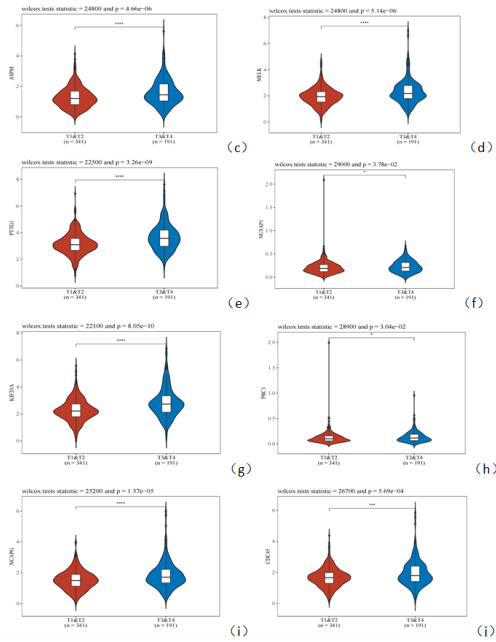
(3A) 最显著的模块来自包含 436 个节点和 772 条边的 PPI 网络。(3B)

图 3 PPI 网络构建及枢纽基因分析

用 STRING 数据库构建 PPI 网络, 在 Cytoscape 软件进行分析, 将 2 个数据集中 915 个 DEGs 导入 STRING 进行 PPI 分析, 综合得分 > 0.7 的纳入 PPI 网络, 最终得到的 PPI 网络共涉及 436 个节点和 772 条边 (图 3A)。cytoHubba 插件选择 MCC 方法排名前 10 的基因 (图 3B)

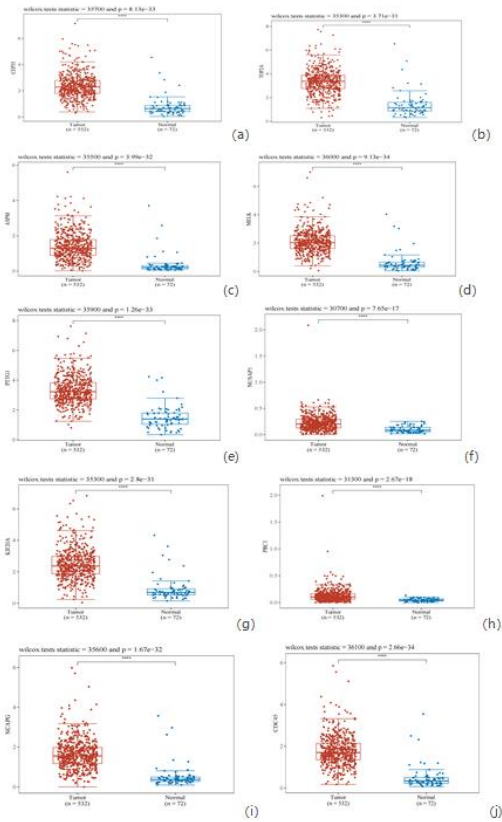
作为 Hub 基因, 包括 CEP55、TOP2A、ASPM、MELK、PTTG1、NUSAP1、KIF20A)、PRC1、NCAPG、CDC45。

2.4 Hub 基因表达与临床病理参数



(a-j) 基因依次为: CEP55、TOP2A、ASPM、MELK、PTTG1、NUSAP1、KIF20A、PRC1、NCAPG、CDC45

图 4 与 T 分类相关的 HCC 患者临床病理参数:

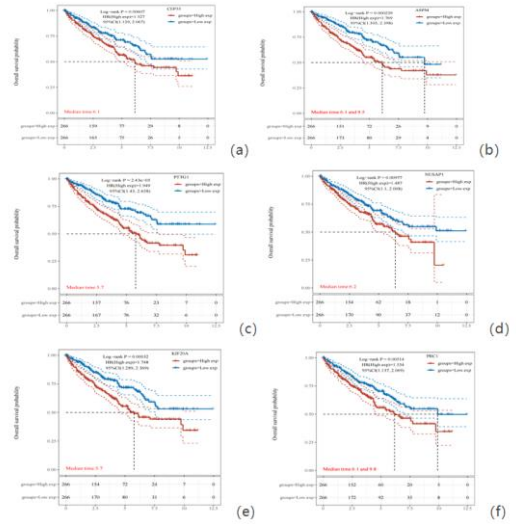


(a-j) 基因依次为: CEP55、TOP2A、ASPM、MELK、PTTG1、NUSAP1、KIF20A、PRC1、NCAPG、CDC45

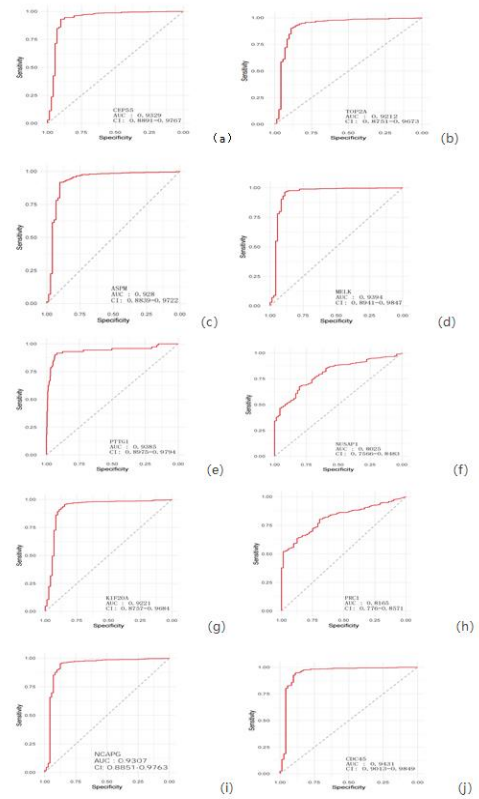
图 5 肿瘤样本和正常样本中 Hub 基因的表达。

在 TCGA 数据库分析 10 个 Hub 基因的表达与 ccRCC 患者临床病理参数的相关性。这些基因的表达与 T 分期相关 ($P < 0.05$) (图 4)。10 个 Hub 基因在癌症样本和癌旁样本中各种基因的表达。(图 5)

2.5 Hub 基因表达与生存分析



(a-f) 依次为 CEP55、ASPM、PTTG1、NUSAP1、KIF20A、PRC1 基因
图 6 10 个中心基因的总生存 (OS) 曲线。HCC 中各种基因高表达和低表达的总生存 (OS) 曲线



(a-j) 基因依次为: CEP55、TOP2A、ASPM、MELK、PTTG1、NUSAP1、KIF20A、PRC1、NCAPG、CDC45

图 7 10 个中心基因的受试者工作特征曲线

使用 R 包 survival 和 survminer 绘制 Kaplan-Meier (KM) 生存曲线, 生存曲线分析结果表明:CEP55 (HR=1.527, P=0.00607)、ASPM (HR=1.769, P=0.00239)、PTTG1 (HR=1.949, P=2.43e-05)、NUSAP1 (HR=1.487, P=0.00977)、KIF20A (HR=1.748, P=0.00032)、PRC1 (HR=1.534, P=0.00514), 这些基因的高表达预示着 ccRCC 患者较差的生存结果 (P<0.05) (图 6), 而 TOP2A、MELK、NCAPG、CDC45 则没有。随后, Hub 基因的 ROC 曲线分析结果显示所有 Hub 基因都能区分 ccRCC 组织与正常肾组织 (图 7)。

3 讨论

本研究从 GEO 和 TCGA 数据库下载 ccRCC 相关数据, 将这两个数据进行联合分析, 彼此验证, 从而得出 915 个共有的 DEGs。接着对这 915 个基因进行功能富集 GO 分析以及 KEGG 分析。而这些生物通路或功能和肿瘤的发生和发展密切相关。再通过构建蛋白质-蛋白质相互作用网络显示出多个高连通性节点, CEP55、TOP2A、ASPM、MELK、PTTG1、NUSAP1、KIF20A、PRC1、NCAPG、CDC45 被发现为 Hub 基因。在 TCGA 数据库进一步的生存分析、ROC 曲线分析选出 6 个核心基因, 包括 CEP55、ASPM、PTTG1、NUSAP1、KIF20A、PRC1。这些基因与 ccRCC 的早期诊断、肿瘤分期和不良预后相关。

中心体蛋白 55 (CEP55), 是一种与细胞分裂和中心体功能密切相关的基因。许多研究表明, CEP55 在多种肿瘤中与增殖、侵袭密切相关^[3]。Chen H 等^[4]研究表明, CEP55 在 ccRCC 中表达升高并通过激活 PI3K/AKT/mTOR 通路促进 ccRCC 细胞上皮间质转化。人类的异常纺锤体微管组装蛋白 (ASPM), 其编码的蛋白质在细胞有丝分裂过程中起着重要作用^[5]。能维持肿瘤细胞的干细胞特性^[6], 对肿瘤的发生发展起到推动作用。垂体肿瘤转化基因 1 (PTTG1), 有研究发现^[7-8], PTTG1 对细胞周期相关蛋白的表达进行调控, 进而促进肾透明细胞癌细胞的增殖。核仁与纺锤体相关蛋白 1 (NUSAP1) 属于一种关键的有丝分裂调节因子。研究表明, 恶性肿瘤的发生发展与 NUSAP1 的异常表达有着紧密联系^[9]。驱动蛋白家族成员 20A (KIF20A) 是驱动蛋白家族中有一个成员, 蛋白过量表达与肿瘤的增殖、侵袭迁移以及耐药等状况有着紧密的联系^[10]。胞质分裂蛋白 1 (PRC1) 是参与胞质分裂调控的关键蛋白^[11-12], PRC1 靶向调控 Hippo 通路促进肾癌细胞增殖和迁移能并与较差的生存预后相关^[13]。

综上所述, 通过生物信息学分析, 我们研究确定了 ccRCC 中的 915 个差异表达基因。差异表达基因揭示了 ccRCC 的机制, 增进了我们对发病机制和预后的理解。通过分析, 确定了 6 个关键基因, 包括 CEP55、ASPM、PTTG1、NUSAP1、KIF20A、PRC1, 这些基因在 ccRCC 的早期诊断、肿瘤分期及不良结局中起着关键作用, 期望能改善 ccRCC 患者的预后和治疗策略。

参考文献:

- [1]Padala SA, Barsouk A, Thandra KC, et al. Epidemiology of Renal Cell Carcinoma. World J Oncol. 2020;11(3):79-87. doi:10.14740/wjon1279
- [2]Escudier B, Porta C, Schmidinger M, et al. Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up[†]. Ann Oncol. 2019;30(5):706-720. doi:10.1093/annonc/mdz056
- [3]Wang G, Chen B, Su Y, et al. CEP55 as a Promising Immune Intervention Marker to Regulate Tumor Progression: A Pan-Cancer Analysis with Experimental Verification[J]. Cells, 2023, 12(20): 2457.
- [4]Chen H, Zhu D, Zheng Z, Cai Y, Chen Z, Xie W. CEP55 promotes epithelial-mesenchymal transition in renal cell carcinoma through PI3K/AKT/mTOR pathway. Clin Transl Oncol. 2019;21(7):939-949. doi:10.1007/s12094-018-02012-8
- [5]Capecchi MR, Pozner A. ASPM regulates symmetric stem cell division by tuning Cyclin E ubiquitination. Nat Commun. 2015;6:8763. Published 2015 Nov 19. doi:10.1038/ncomms9763
- [6]吴莹莹,金木兰.ASPM 在肿瘤中的作用机制及最新研究进展[J].诊断病理学杂志,2021,28(02):128-130.
- [7]宋章兴,崔应东,李时军,等.肾透明细胞癌病理分级相关基因的 WGCNA 筛选[J].现代泌尿生殖肿瘤杂志,2021,13(01):9-15.
- [8]魏灿,张艳斌,杨晓亮,等.垂体肿瘤转化基因 1 在肾透明细胞癌中的表达及其与预后的关系[J].中华实验外科杂志,2018,35(9):1610-1612.
- [9]Iyer J, Moghe S, Furukawa M, Tsai MY. What's Nu(SAP) in mitosis and cancer?. Cell Signal. 2011;23(6):991-998. doi:10.1016/j.cellsig.2010.11.006
- [10]苏天宇,张有为,王翔.KIF20A 在肿瘤中的作用及调控机制研究[J].徐州医科大学学报,2024,44(05):368-374.

- [11]Li J, Dallmayer M, Kirchner T, Musa J, Gr ü newald TGP. PRC1: Linking Cytokinesis, Chromosomal Instability, and Cancer Evolution. Trends Cancer. 2018;4(1):59–73. doi:10.1016/j.trecan.2017.11.002
- [12]Li S, Motiño O, Lambertucci F, Martins I, Sun L, Kroemer G. Protein regulator of cytokinesis 1: a potential oncogenic driver. Mol Cancer. 2023;22(1):128. Published 2023 Aug 10. doi:10.1186/s12943-023-01802-1
- [13]万梓宇,王霞,郑张节,等.PRC1 靶向调控 Hippo 通路促

进肾癌细胞增殖和迁移能力[J].武汉大学学报(医学版),2025,46(01):36–40.DOI:10.14188/j.1671-8852.2024.0087.

作者简介: 高风霞 (1989—), 女, 汉, 硕士, 分子诊断相关研究。通讯作者: 高慧洁 (1988—), 女, 汉, 硕士, 疾病检验诊断相关研究。

基金项目: 河北省高等学校科学技术研究项目 (BJK2024152)。