

整合多个基因表达谱数据识别骨肉瘤发生和进展相关的关键基因

韩小玲,王胜纯,韩俏英,梁 兰* (广东医科大学,广东东莞 523808)

摘要:目的 了解骨肉瘤发生和进展相关的关键基因。方法 将GEO数据库中多个骨肉瘤基因表达谱整合为一个表达矩阵,对骨肉瘤关键基因进行功能注释和通路富集分析,识别模块内与骨肉瘤发生与发展密切相关的关键基因。结果 4个骨肉瘤数据集整合成含有98个样本、13 914个基因表达谱的矩阵文件,识别出骨肉瘤相关程度最高的turquoise模块($\text{cor}=0.71$, $P=4\text{e-}16$),包括LDHB、CBX2、CCT3、KPNA2、PPAT、MSH6、UBE2C、SNRPA1、CCDC58、TMEM97和KIF2C等11个骨肉瘤关键基因,其中SNRPA1、CCDC58、TMEM97、KIF2C和UBE2C未见报道。结论 发现11个关键基因与骨肉瘤发生和进展相关。

关键词:骨肉瘤; GEO数据库; 关键基因

中图分类号: R 730.231

文献标志码: A

文章编号: 2096-3610 (2023) 06-0613-06

Identification of key genes for osteosarcoma development and progression using multiple gene expression profile

HAN Xiao-ling, WANG Sheng-chun, HAN Qiao-ying, LIANG Lan* (Guangdong Medical University, Dongguan 523808, China)

Abstract: Objective To identify key genes for osteosarcoma occurrence and progression using multiple gene expression profile. Methods Multiple gene expression profile of osteosarcoma in GEO database was integrated into an expression matrix. The key osteosarcoma-associated genes underwent functional annotation and pathway enrichment analysis, followed by identifying the key genes for osteosarcoma development and progression. Results Four osteosarcoma datasets were integrated into a matrix file containing 98 samples and 13 914 gene expression profiles. The turquoise module was the most associated module for osteosarcoma ($\text{cor}=0.71$, $P=4\text{e-}16$), and included LDHB, CBX2, CCT3, KPNA2, PPAT, MSH6, UBE2C, SNRPA1, CCDC58, TMEM97 and KIF2C, of which SNRPA1, CCDC58, TMEM97, KIF2C, and UBE2C have not been reported. Conclusion Eleven key genes are associated with osteosarcoma occurrence and progression.

Key words: osteosarcoma; GEO database; key genes

骨肉瘤是一种常见的起源于间充质组织的骨肿瘤疾病,好发于10~20岁年龄段的青少年人群及老年人群。由于骨肉瘤具有恶性程度高、易转移的特点,患者预后总体较差,相关数据显示对于非转移型骨肉瘤患者而言,其5年生存率可达到60%~70%,而对于转移性骨肉瘤患者,5年生存率仅为20%~30%。深入研究并阐明骨肉瘤的发生、发展及转移分子机制是有效改善患者预后的必要前提,也是目前骨肉瘤研究领域的热点。本研究通过整合多个骨肉瘤基因表达谱数据集,并基于加权基因共表达网络的分析方法识别骨肉瘤进展密切相关的基因,从而试图为开展骨肉瘤分子机制

研究提供新的且重要的候选靶点分子。

1 材料和方法

1.1 数据获取

借助“GEOquery”R包^[1]分别下载和整理GEO数据库中骨肉瘤相关数据集GSE28424、GSE36001、GSE87624和GSE126209,处理过程主要包括基因注释、去重、表达谱进行标准化和取log值等步骤。利用“sva”包中的combat函数合并上述4个骨肉瘤数据集标准化后的表达谱并去除批次效应^[2]。合并后的骨肉瘤基因表达矩阵用于后续加权基因共表达网络构建和

收稿日期: 2023-05-18

基金项目: 国家自然科学基金项目(81702953),广东省医学科研基金项目(A2023169)

作者简介: 韩小玲(1982-),女,本科,实验师, E-mail: 2359713979@qq.com

通信作者: 梁 兰(1986-),女,硕士研究生,实验师, E-mail: 610194795@qq.com

分析。另外,通过“GEOquery” R包下载含有骨肉瘤转移样本与非转移样本信息的数据集 GSE21257,以及基于 UCSC xena 网站 (<https://xenabrowser.net/datapages/>) 下载含有患者预后信息的 Target 数据库中的骨肉瘤数据集分析骨肉瘤关键基因。

1.2 WGCNA

本研究中 WGCNA 过程主要是基于“WGCNA” R 包实现^[3],具体包括基因过滤、离群样本删除、软阈值(β)筛选、TOM 矩阵计算、模块合并及相关性计算几个关键步骤。WGCNA 运行过程中涉及的几个重要参数分别进行设置,其中 minModuleSize 设置为 30,而 MEDissThres 设置为 0.25。最后根据样本信息与模块 ME 相关性程度及 P 值确定骨肉瘤关键基因模块并用于后续骨肉瘤关键基因的筛选,骨肉瘤关键基因的筛选条件为: $\text{cor.gene GS} > 0.6$, $\text{cor.gene MM} > 0.7$ 。

1.3 功能注释和通路分析

为阐明骨肉瘤发生和进展过程中参与的生物学功能及通路途径,本文对骨肉瘤关键模块内基因进行了 GO 生物学功能注释和 KEGG 通路分析,该分析主要是基于 R 中“clusterProfiler”包^[4]实现。

1.4 骨肉瘤关键基因的分析

对 WGCNA 识别出的关键基因,通过 NCBI pubmed 网站探究基因在骨肉瘤及其他肿瘤中的研究情况并基于 Cytoscape 软件中的“GeneMANIA”插件分析关键基因及其相关基因共表达情况^[5];通过“corrplot”包 (<https://github.com/taiyun/corrplot>) 分析和可视化关键基因相关性以及部分关键基因与细胞增殖标记基因、细胞周期标记基因和转移标记基因的相关性;利用 R 中“pROC”包对关键基因诊断性能进行分析并绘制受试者工作特征曲线(ROC 曲线)和计算 AUC 值^[6];再者,基于 GSE21257 数据集分析关键基因是否与转移相关;利用 Target 数据库中的骨肉瘤样本进行上述基因的预后性能分析。

1.5 统计学处理

本研究的数据分析及可视化过程均基于 R3.6.3 和 Rstudio 软件完成,两组间基因表达差异分析利用 t 检验。预后分析采用 Kaplan-Meier 方法,并根据基因表达值的最佳值将样本分为高表达组和低表达组,组间生存差异采用 logrank 检验进行。以 $P < 0.05$ 为差异有统计学意义。

2 结果

2.1 数据获取

本研究共用到 6 个骨肉瘤相关数据集,其中数据

集 GSE28424、GSE36001、GSE87624 和 GSE126209 均包含有正常样本和骨肉瘤样本,该 4 个数据集经整合后形成具有 98 个样本、13 914 个基因表达谱的矩阵文件,合并后的矩阵用于 WGCNA 与诊断性能分析;数据集 GSE21257 为取材于化疗前骨肉瘤样本,共含有骨肉瘤转移样本 34 例,非转移性骨肉瘤样本 19 例,来源于 Target 数据库含有患者随访信息的骨肉瘤样本 87 例,数据集 GSE21257 和 Target 中的基因表达谱数据用于骨肉瘤关键基因转移与预后的相关分析。

2.2 加权基因共表达网络分析

考虑到后续分析过程中的运算负荷,本文选取了基因表达谱中方差变化最大的前 25% 的基因用于加权基因共表达网络构建和分析,图 1 描述基于基因表达谱和样本特征所构建的聚类树。软阈值(β 值)筛选是构建加权基因共表达网络的关键环节,按照图 2A 所示,本文选取软阈值为 6 用于随后的拓扑矩阵的计算,而图 2B 所示结果也说明本研究中构建网络达到无尺度网络标准。随后,拓扑重叠矩阵计算完成后,按照两两基因表达相异性计算公式,具有共表达关系的基因被分配到同一模块内,图 2C 中分层聚类树展示了 3 479 个基因被分配到了不同的模块内。最后通过对模块与样本特征相关性分析(图 2D),发现 turquoise 模块与骨肉瘤呈非常强的正相关,本文初步判断该模块内基因可能在骨肉瘤的发生和进展过程中发挥重要作用。因此, turquoise 模块作为骨肉瘤关键基因模块被用于后续的分析。

2.3 功能注释和通路富集分析

为明确 Turquoise 模块内基因与骨肉瘤发生、发展的相关性,本文对该模块内的 1 034 个基因进行了 GO 生物学过程(BP)注释及 KEGG 通路富集分析,见图 3。其中, BP 分析显示 Turquoise 模块内基因显著富集在 DNA 复制、核分裂、线粒体分裂及细胞周期调控相关的生物学过程(图 3A);而 KEGG 通路富集分析结果则显示该模块内基因主要富集在细胞周期、DNA 复制、P53 信号通路、TGF- β 信号通路及蛋白消化吸收等通路(图 3B)。

2.4 骨肉瘤关键基因的识别

在确定 Turquoise 模块内基因与骨肉瘤发生、发展密切关系后,按照 $\text{cor.gene GS} > 0.6$, $\text{cor.gene MM} > 0.7$ 的过滤条件本文在该模块内识别出 11 个骨肉瘤的关键基因,包括 LDHB、CBX2、CCT3、KPNA2、PPAT、MSH6、UBE2C、SNRPA1、CCDC58、TMEM97 和 KIF2C(见图 4)。经文献检索本文还发现已有多篇

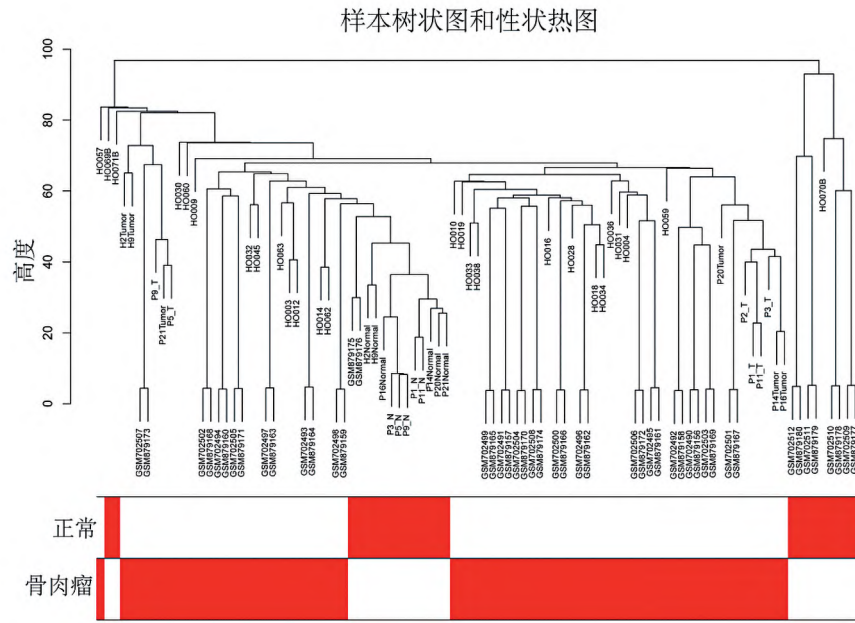


图1 方差变化最大的前 25% 基因表达谱构建的树状图和样本特征热图

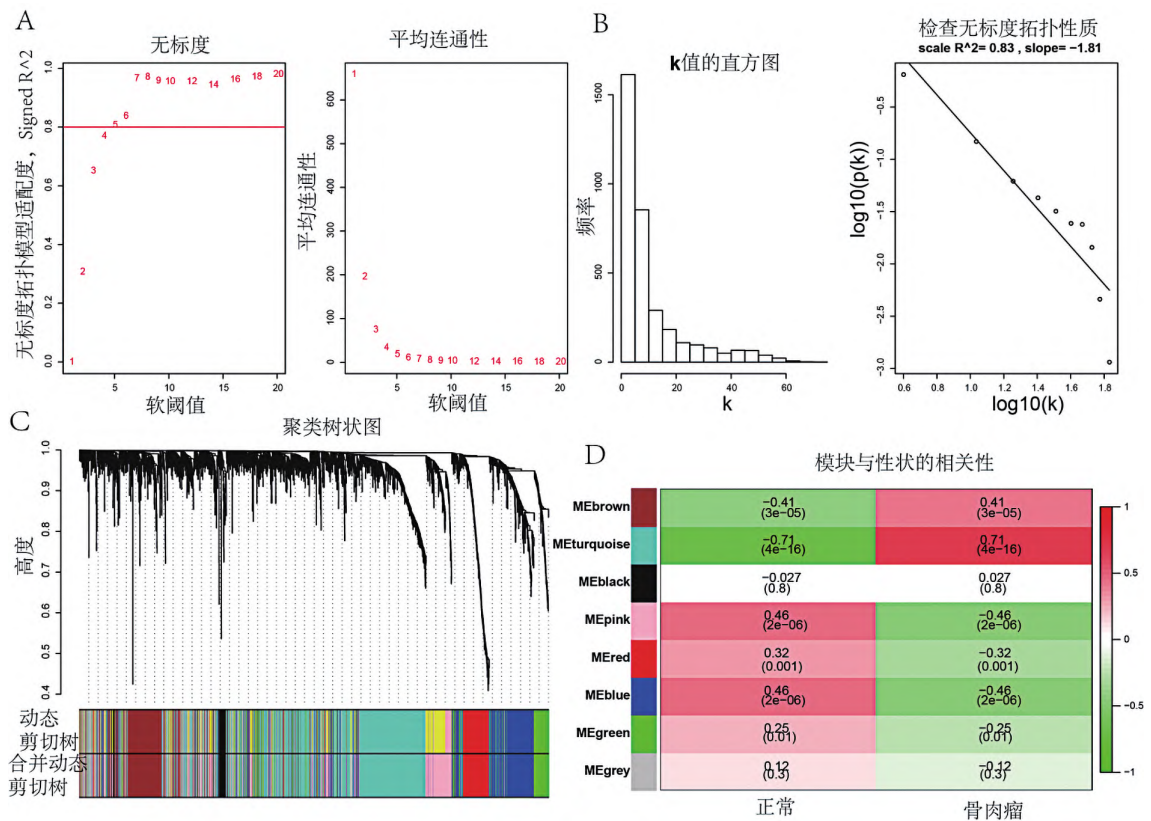


图2 加权基因共表达网络的构建和分析

关于 LDHB、CBX2、CCT3、KPNA2、PPAT 和 MSH6 在骨肉瘤中的研究报道,而暂未见关于其他 5 个基因 SNRPA1、CCDC58、TMEM97、KIF2C 和 UBE2C 与骨肉瘤研究的报道。

2.5 骨肉瘤关键基因分析

相关性分析显示上述 11 个骨肉瘤相关基因存在极强的相关性(图 5A),而利用“GeneMANIA”插件分析关键基因及其相关基因共表达情况同样显示 11 个骨

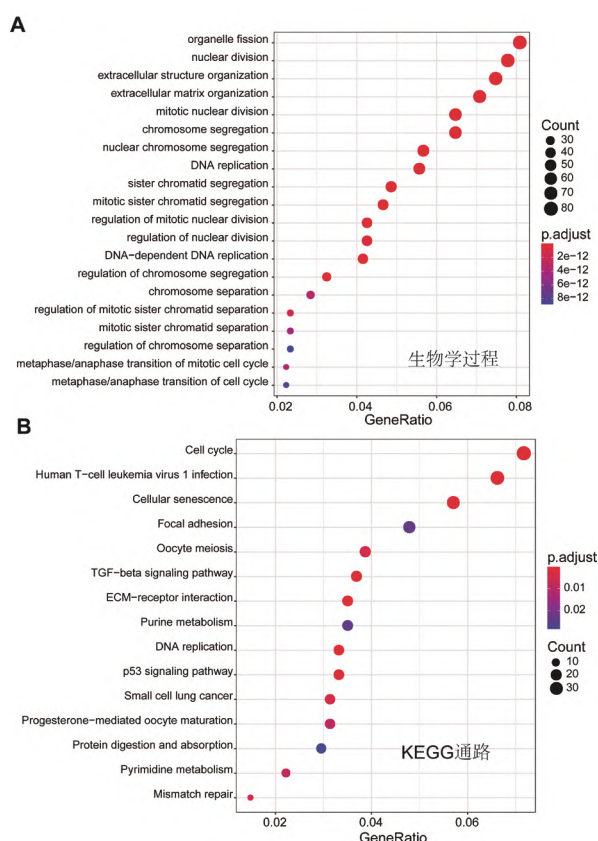


图 3 turquoise 模块内基因生物学功能和通路富集分析

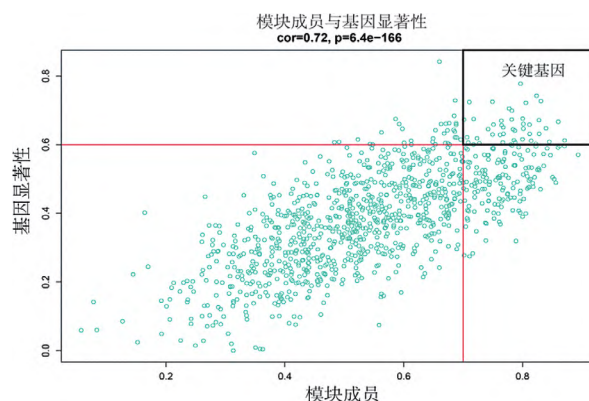


图 4 turquoise 模块内骨肉瘤关键基因的识别

肉瘤关键基因之间联系紧密(图 5B)。随后本文进行了 SNRPA1、CCDC58、TMEM97、KIF2C 和 UBE2C, 与增殖标记基因 (MKI67、PCNA) 相关性分析, 结果显示他们具有显著正相关(图 5C), 而与细胞周期促进因子 (CDK3、CCNE2、CCNB1、CCNA2 和 CCNB2) 也同样具有显著正相关, 与细胞周期抑制因子 (CDKN1C、CDKN2B) 则呈负相关(图 5D)。考虑到转移对骨肉瘤预后产生的不利影响, 本文还对 SNRPA1、CCDC58、TMEM97、KIF2C 和 UBE2C 与转移标记基因 CDH2 进行相关性分析, 结果也显示具有显著的正相关(图 5E)。

2.6 骨肉瘤关键基因诊断性能和预后性能分析

在对 11 个骨肉瘤关键基因的诊断性能进行评

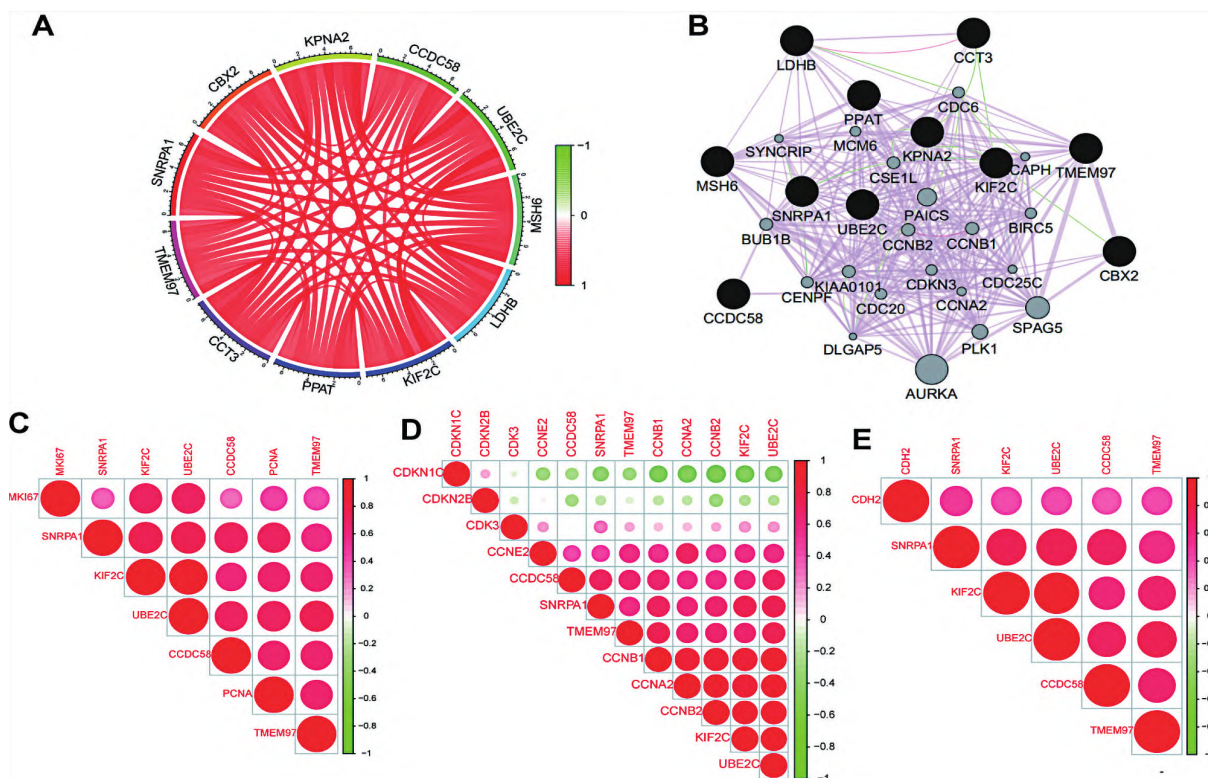


图 5 11 个骨肉瘤关键基因的相关性和共表达分析识别

估, 本文发现在用于诊断骨肉瘤方面 11 个基因均呈现良好的诊断性能 (AUC 值 >0.8 , 图 6A)。本文利用数据集 GSE21257 分析了 11 个骨肉瘤关键基因在骨肉瘤转移组和非转移组表达是否具有差异性, 结果显示 CCDC58、KPNA2、LDHB、MSH6 和 UBE2C 在骨肉瘤转移组显著上调 (图 6B), 且 CCDC58、KPNA2、LDHB、MSH6 和 UBE2C 判断骨肉瘤转移方面具有一定的诊断性能 (AUC 值 >0.7), 见图 6C。另外 KM 分析还显示 11 个基因中 CCDC58 与骨肉瘤不良预后显著相关, 是影响骨肉瘤预后的风险因子 ($P < 0.05$, 图 6D)

3 讨论

作为一种临床上常见的骨科肿瘤疾病, 深入研究骨肉瘤机理十分有必要的, 而基于基因表达谱, 寻找疾病发生、发展过程中的关键基因并探寻其调控关系是目前进行肿瘤等复杂疾病机制研究的有效研究方法。在骨肉瘤关键基因识别方面的相关研究目前也有报道, 如学者林文韬等^[7]基于骨肉瘤数据集基因表达谱识别出 GMFG 促进了骨肉瘤进展, 是骨肉瘤关键基因。冯晓飞^[8]也利用转录组数据分析并识别出 CDC6、FEN1、OIP5、GINS2、RFC3、TRIP13、MCM10、

AURKA、TIPIN 和 DTYMK 10 个骨肉瘤关键基因, 这为骨肉瘤机制研究提供了重要的候选分子。但对于这些骨肉瘤相关研究, 本文发现存在弊端, 体现在一方面研究者往往仅依靠单个数据集, 而这样识别出的关键基因往往缺乏普遍性; 另一方面在识别关键基因方法方面仅依靠 PPI 网络, 很难鉴别出新的关键基因。基于此, 本文整合了多个骨肉瘤数据集, 并利用加权基因共表达网络分析方法进行骨肉瘤潜在关键基因的识别, 从而为骨肉瘤机制研究提供有效候选分子。

在有效去除 4 个骨肉瘤数据集间的批次效应并将其基因表达谱合并后, 本文选用方差变化最大的 3 479 个基因构建 WGCNA 网络并分析识别出与骨肉瘤相关性最强的 turquoise 模块, 功能注释和通路分析显示该模块内基因涉及到细胞周期、DNA 复制、P53 信号通路、TGF- β 信号通路及蛋白消化吸收等人体调控过程, 而相关研究已表明上述功能和通路参与了包括骨肉瘤在内多种肿瘤的进程^[9], 初步说明了本文分析结果的正确性。按照严格的筛选条件, 本文共识别出 11 个与骨肉瘤发展密切相关的基因, 对于其中的 7 个基因: 已有报道 LDHB 是骨肉瘤不良预后预测因子^[10]; MiR-141-3p 降低 LDHB 表达水平能有

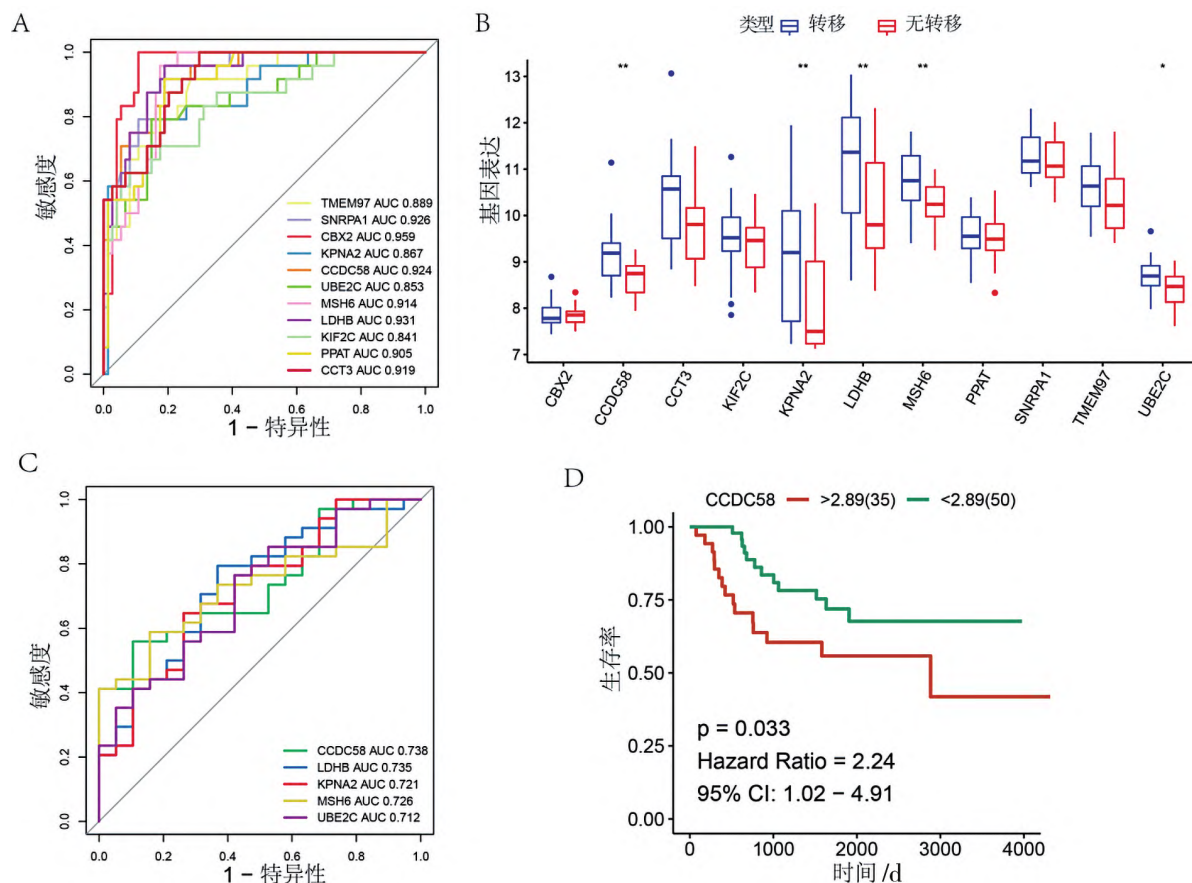


图6 骨肉瘤关键基因诊断与预后性能分析

效遏制骨肉瘤恶性程度^[11]; CBX2 可以作为 miR- let-7a 的靶标促进骨肉瘤进展^[12]; CCT3 也被认为是骨肉瘤重要的候选驱动基因^[13]; KPNA2 在骨肉瘤中表达显著上调并具有作为骨肉瘤诊断标记物的潜能^[14]; PPAT 也被认为是骨肉瘤潜在的治疗靶点^[15]; 另外, MSH6 在骨肉瘤中的研究也较多, 有报道 CSE1L 与 MSH6 相关作用促进了骨肉瘤进展并能有效预测患者不良预后^[16]。对于剩余 5 个骨肉瘤相关关键基因 SNRPA1、CCDC58、TMEM97、KIF2C 与 UBE2C 在肝癌、肺癌和结直肠癌等恶性肿瘤中表达已有相关报道, 但关于他们在骨肉瘤中的研究尚少。本研究中, 通过共表达和相关性分析初步认识到 11 个基因间存在非常强的共表达强度与相关性, 这也暗示着对于在骨肉瘤中研究尚未涉及的 5 个关键基因确实也在骨肉瘤的发生进展中起到重要作用。而随后进行的 5 个关键基因与肿瘤细胞增殖标记基因、细胞周期标记基因的相关性分析也进一步证明了本文的假设。转移的发生严重影响了骨肉瘤患者的结局, 导致了患者的不良预后, 本研究对识别出的 11 个骨肉瘤关键基因与转移相关性进行分析, 结果显示多个基因在转移样本中显著上调, 且与转移标记基因 CDH2 呈正相关, 提示他们在骨肉瘤转移进展中也发挥重要作用。随后本文还对 11 个骨肉瘤关键基因进行诊断性分析, 11 个基因在用于判断骨肉瘤方面均展示良好的诊断性能, 提示 11 个关键基因在骨肉瘤中的重要的应用价值, 而对于基因 CCDC58, 不仅具有良好的预后评估性能, 其在判断骨肉瘤转移与否方面也显示了潜在价值。

综上所述, 本研究通过整合多个骨肉瘤数据集基因表达谱, 利用先进的加权基因共表达网络分析方法共识别出 11 个骨肉瘤相关的关键基因, 其中有 5 个在骨肉瘤相关性研究中还未涉及, 这为本文进行骨肉瘤分子机制的研究以及诊断标记物的筛选提供了重要的候选分子, 而对于其中的 CCDC58 所展现的良好预后和转移诊断标记物的潜在价值也值得进一步深入。

参考文献:

- [1] MELTZER D P S. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor[J]. *Bioinformatics*, 2007, 23(14): 1846-1847.
- [2] LEEK J T, JOHNSON W E, PARKER H S, et al. The sva package for removing batch effects and other unwanted variation in high-throughput experiments[J]. *Bioinformatics*, 2012, 28(6): 882-883.
- [3] LANGFELDER P, HORVATH S. WGCNA: An R package for weighted correlation network analysis[J]. *BMC Bioinformatics*, 2008, 9(1): 559.
- [4] YU G, WANG L G, HAN Y, et al. ClusterProfiler: An R package for comparing biological themes among gene clusters[J]. *Omics*, 2012, 16(5): 284-287.
- [5] MONTOJO J, ZUBERI K, RODRIGUEZ H, et al. GeneMANIA: Fast gene network construction and function prediction for Cytoscape[J]. *F1000Res*, 2014, 3: 153.
- [6] ROBIN X, TURCK N, HAINARD A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves[J]. *BMC Bioinformatics*, 2011, 12(1): 77.
- [7] 林文韬, 王武炼, 肖莉莉, 等. 基于 GEO 数据库发现骨肉瘤关键基因 GMFG 及其生物学功能[J]. *福建医药杂志*, 2020, 42(2): 112-115.
- [8] 冯晓飞. 骨肉瘤关键候选基因和通路分析及构建预测总生存期的诺模图[D]. 兰州: 兰州大学, 2020.
- [9] 吴进, 刘庆军, 陈志达, 等. p38 MAPK/p53 信号通路调控骨肉瘤细胞中 *Ether a go-go* 表达的研究[J]. *中国癌症防治杂志*, 2015, 7(5): 325-329.
- [10] LI C, CHEN Y, BAI P, et al. LDHB may be a significant predictor of poor prognosis in osteosarcoma[J]. *Am J Transl Res*, 2016, 8(11): 4831-4843.
- [11] WANG L. MiR-141-3p overexpression suppresses the malignancy of osteosarcoma by targeting FUS to degrade LDHB[J]. *Biosci Rep*, 2020, 40(6): BSR20193404.
- [12] HAN Q, LI C, CAO Y, et al. CBX2 is a functional target of miRNA let-7a and acts as a tumor promoter in osteosarcoma[J]. *Cancer Med*, 2019, 8(8): 3981-3991.
- [13] XIONG Y, WU S, DU Q, et al. Integrated analysis of gene expression and genomic aberration data in osteosarcoma (OS)[J]. *Cancer Gene Ther*, 2015, 22(11): 524-529.
- [14] JIANG L, LIU J, WEI Q, et al. KPNA2 expression is a potential marker for differential diagnosis between osteosarcomas and other malignant bone tumor mimics[J]. *Diagn Pathol*, 2020, 15(1): 135.
- [15] CHAIYAWAT P, SETTAKORN J, SANGSIN A, et al. Exploring targeted therapy of osteosarcoma using proteomics data[J]. *Onco Targets Ther*, 2017, 10: 565-577.
- [16] CHENG D D, LIN H C, LI S J, et al. CSE1L interaction with MSH6 promotes osteosarcoma progression and predicts poor patient survival[J]. *Sci Rep*, 2017, 7: 46238.