



论文

大肠癌患者生存相关基因的筛选与网络调控分析

齐鲁^①, 丁彦青^{①②*}

① 南方医科大学基础医学院病理学系, 广州 510515;

② 南方医科大学南方医院病理科, 广州 510515

* 联系人, E-mail: dyq@fimmu.com

收稿日期: 2013-10-10; 接受日期: 2013-11-04; 网络版发表日期: 2014-03-10

国家自然科学基金联合基金(批准号: U1201226)和国家自然科学基金(批准号: 30670967)资助项目

摘要 本研究筛选了与大肠癌患者生存期相关的关键基因, 以探索这些基因所参与的信号调控网络. 在前期研究中, 通过对大肠癌相关表达谱数据 GSE17538 使用显著性分析软件 SAM3.01 对大肠癌患者生存期相关的基因进行筛选, 得到了与大肠癌患者生存期相关的基因 235 个. 对这 235 个基因进行以下筛选和分析: 首先对 235 个基因进行转录因子结合位点富集分析, 筛选含有转录因子结合位点数目大于 7 个的基因, 再将这些基因与大肠癌上调基因集取交集, 最后将筛选出的基因进行生存分析及网络调控分析, 明确这些基因与大肠癌患者生存期的关系及在大肠癌细胞信号调控网络中所参与的网络. 通过上述分析筛选出的与大肠癌患者生存期相关, 受转录因子调控较多, 且在大肠癌中表达上调的基因有 6 个, 分别为 *STX2*, *PODXL*, *KLK6*, *GRB10*, *EHBP1* 和 *CREB5*. 这些基因所参与的网络调控网络与大肠癌转移相关信号通路相关. 大肠癌患者生存期与大肠癌转移密切相关, 生存期相关的 6 个基因通过调控大肠癌转移相关信号通路影响患者的生存期.

关键词生存期
大肠癌
转录因子
生物信息学

大肠癌患者患癌后的生存时间受多方面的影响, 实际研究过程中, 大肠癌细胞中部分基因的表达变化影响着大肠癌患者的生存期及预后, 因此, 阐明大肠癌患者患癌后生存时间变化的分子机制, 对指导大肠癌的治疗及评估预后有着重要意义. 在前期研究中, 通过对 NCBI(<http://www.ncbi.nlm.nih.gov/>)的 GEO(gene expression omnibus)数据库中表达谱数据 GSE17538 使用显著性分析软件 SAM3.01 对大肠癌患者生存时间相关基因进行筛选, 能够得到与大肠癌患者生存时间相关的基因 235 个, 并且对这 235 个基因进行 KEGG(kyoto encyclopedia of genes and gen-

omes)通路富集分析, 发现这些基因与大肠癌转移事件密切相关^[1]. 本研究主要对这 235 个与大肠癌患者生存时间相关的基因进行进一步分析, 筛选这些基因中较为重要的基因, 通过对这些基因所编码蛋白的相互作用网络调控情况分析, 阐明这些基因在影响大肠癌患者生存时间变化过程中可能起到的作用.

1 材料与方法

首先使用 TFM-Explorer(the transcription factor matrix explorer)工具^[2]对上述 235 个基因的转录调控

引用格式: 齐鲁, 丁彦青. 大肠癌患者生存相关基因的筛选与网络调控分析. 中国科学: 生命科学, 2014, 44: 481-487

英文版见: Qi L, Ding Y Q. Screening and regulatory network analysis of survival-related genes of patients with colorectal cancer. Sci China Life Sci, 2014, 57: in press

区域内转录因子结合位点(transcription factor binding site, TFBS)进行富集分析. 235 个基因中, 部分转录因子结合位点同时存在于多个基因中, 说明这类转录因子调控的生存相关基因较多, 通过得分排序, 筛选得到这类转录因子结合位点的集合. 同样, 许多基因的转录调控区域同时含有上述多个转录因子结合位点, 筛选含有上述转录因子结合位点数目大于 7 个的基因.

由于这些基因在大肠癌中表达情况未知, 若这些基因在大肠癌中表达上调则对患者生存时间更有意义. 因此需要使用在大肠癌中表达上调基因集对上述筛选出的基因进行过滤. 在此使用 GEO 数据库中的 GSE21815 和 GSE21510 筛选大肠癌表达上调的基因集. 由于之前的分析得知 235 个生存时间相关的基因富集结果与大肠癌转移事件相关, 因此选择 GSE21815 数据中的 9 例正常大肠数据及 49 例处于淋巴结转移期的大肠癌数据进行比较筛选. 同样对 GSE21510 数据中的 25 例正常大肠数据和 39 例处于淋巴结转移期的大肠癌数据进行比较筛选. 筛选差异表达基因使用工具为 Genespring 11.5, 选择 Benjamini Hochberg 的方法评估错误发现率, 调整筛选基因的 P 值小于 0.01, 并且选择大肠癌比正常大肠基因表达变化大于 2 倍的基因.

将此前筛选出的具有转录因子结合位点数目大于 7 个的基因与上两组表达谱筛选出的上调差异表达基因分别取交集, 然后将过滤后的两组基因再取交集(图 1), 将最终筛选出的基因使用 STRING 工具^[3]进行网络调控情况分析, 再提取网络中的蛋白节点通过 KEGG 数据库进行通路富集分析, 明确网络中的蛋白所参与的分子事件.

2 结果

2.1 生存相关基因转录因子结合位点富集结果

为了明确 235 个生存相关基因受转录因子的调控情况, 使用 TFM-Explorer 工具对 235 个基因转录调控区域的转录因子结合位点进行富集分析, 筛选得到 25 个在上述 235 个基因中出现频率较高的转录因子结合位点, 且转录起始位点上游 20~240 序列间的转录因子结合位点最多. 富集的 25 个转录因子结合位点中, *SOX2*, *NFIC* 和 *TBP* 基于定位基序的不同

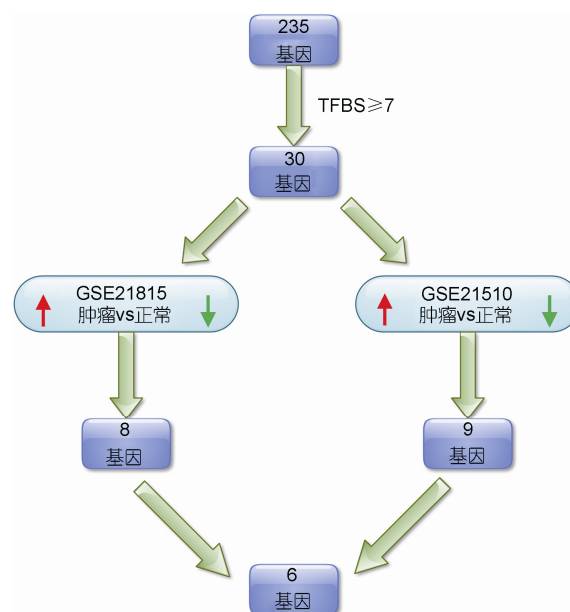


图 1 生存相关关键基因筛选流程

均出现了 1 次重复, 因此共有 22 类转录因子结合位点, 这 22 类转录因子分别为 *SP1*, *EGR1*, *ELK4*, *EN1*, *ESRRB*, *ETS1*, *FOXC1*, *FOXD1*, *FOXQ1*, *GATA2*, *KLF4*, *NFIC*, *NFYA*, *PDX1*, *RELA*, *RREB1*, *SOX2*, *SOX5*, *SPIB*, *SRF*, *T* 和 *TBP*(表 1). 这些转录因子可能分别调控着与生存相关的 235 个基因中的相应基因, 所以这些转录因子在调控大肠癌患者生存时间上有着重要意义.

2.2 生存相关关键基因的筛选结果

由于不同基因受到转录调控情况不一致, 基因上游转录因子结合位点的数量也就不一致, 转录因子结合位点越多, 说明此基因受到的转录因子调控越多, 调控也就越复杂, 基因对生命活动的影响就越重要. 多个转录因子结合位点通常组成顺式调控模块(*cis-regulatory module*, CRM), 而有研究表明, 顺式调控模块富集的区域与生长发育相关的转录因子存在着关联, 而与生长发育相关的转录因子在肿瘤的发生发展中同样起着重要作用^[4]. 因此含有上述 25 个转录因子结合位点越多的基因, 在大肠癌发展过程中可能就越重要, 在调控大肠癌患者的生存时间上可能也就越重要. 通过筛选, 含有上述转录因子最多的生存相关基因为 *CREB5* 和 *LOC401317*, 这两个基因含有 25 个转录因子结合位点中的 13 个结合位

表 1 22 类转录因子结合位点对应基因名称

基因符号	基因名称
<i>SPI</i>	Sp1 转录因子
<i>EGR1</i>	早期生长反应因子 1
<i>ELK4</i>	ETS-结构域蛋白(SRF 辅助蛋白 1)
<i>EN1</i>	Engrailed 同源盒蛋白 1
<i>ESRRB</i>	雌激素相关受体 β
<i>ETS1</i>	V-Ets 骨髓成红细胞增多症病毒 E26 癌基因同源物 1
<i>FOXC1</i>	叉头框蛋白 C1
<i>FOXD1</i>	叉头框蛋白 D1
<i>FOXQ1</i>	叉头框蛋白 Q1
<i>GATA2</i>	GATA 结合蛋白 2
<i>KLF4</i>	Kruppel 样转录因子 4
<i>NFIC</i>	核因子 1/C(CCAAT-结合转录因子)
<i>NFYA</i>	核转录因子 $Y\alpha$
<i>PDX1</i>	胰十二指肠同源异型盒蛋白 1
<i>RELA</i>	V-Rel 网状内皮增生病毒癌基因同源物 A
<i>RREB1</i>	Ras 反应元件结合蛋白 1
<i>SOX2</i>	性别决定区 Y 框蛋白 2
<i>SOX5</i>	性别决定区 Y 框蛋白 5
<i>SPIB</i>	Spi-B 转录因子(Spi-1/PU.1 相关)
<i>SRF</i>	血清反应因子
<i>T</i>	T-鼠短尾突变体表型同系物
<i>TBP</i>	TATA 盒结合蛋白

点. 同理含有 12 个转录因子结合位点的基因有 *ETV1*, *MEIS2* 和 *PALLD*, 含有 11 个转录因子结合位点的基因有 *EHBPI*, *MITF* 和 *TNIK*, 含有 10 个转录因子结合位点的基因有 *TGFBIII*. 综合考虑筛选的基因数量以及转录因子数量, 选择转录因子结合位点数大于 7 个的基因有 30 个(表 2). 由于这些基因在大肠癌中表达情况未知, 若这些基因在大肠癌中表达上调则对患者生存时间更有意义. 因为在肿瘤中表达上调的基因通常对肿瘤的发生发展有着积极影响, 若上调基因与患者生存时间呈负相关, 则随着基因的表达上调患者的生存率会出现显著下降, 因此可以根据这些基因的表达情况, 评估患者的生存时间. 其次这些基因为阳性表达, 其在大肠癌组织中表达升高, 而在正常大肠组织中表达降低, 这样更有利于基因在临床诊断中的应用. 并且在细胞信号转导过程中, 表达上调的基因能够将上调信号进行级联传递, 进而影响细胞的生物学行为. 例如 GSEA(gene set enrichment analysis)^[5]等富集工具均使用表达上调基因对不同生物学表型进行富集分析的. 因此本研究需要对转录因子结合位点数大于 7 个的基因根据表达情况进行进一步的筛选. 在此使用 GEO 数据库中

的 GSE21815 和 GSE21510 数据分别筛选大肠癌表达上调的基因集, 然后使用此基因集与上述 30 个基因取交集, 其中与数据 GSE21815 上调基因集取交集后得到的基因数有 8 个, 分别为 *TGFBIII*, *STX2*, *PODXL*, *KLK6*, *GRB10*, *EHBPI*, *CREB5* 和 *ARHGAP5*. 与数据 GSE21510 上调基因集取交集后得到的基因数有 9 个, 分别为 *ZNF83*, *STX2*, *PODXL*, *KLK6*, *GRB10*, *ETV1*, *CREB5*, *CALD1* 和 *EHBPI*. 两组交集基因再取交集, 最终筛选出与患者生存时间相关, 含上述转录因子结合位点较多, 且在大肠癌中表达上调的基因有 6 个, 分别为 *STX2*, *PODXL*, *KLK6*, *GRB10*, *EHBPI* 和 *CREB5*.

2.3 生存相关关键基因的生存分析结果

为了进一步明确上述 6 个基因的生存曲线情况, 使用 GEO 数据库^[6]中的表达谱数据 GSE17536 对上述基因进行生存分析. 数据 GSE17536 与前期研究分析中使用的数据 GSE17538 为包含关系. 其中 GSE17536 数据包含 177 例具有患者生存时间信息的大肠癌组织表达谱数据, 其芯片平台为 Affymetrix Human Genome U133 Plus 2.0 Array. 从此表达谱数据中提取上述 6 个基因的 177 例表达值. 分别将 6 个基因的表达值从小到大进行排序并分为低、中、高 3 组, 这样每组包含 59 例病例的表达值, 且 3 组表达值是递增关系. 然后使用 Kaplan-Meier 生存分析的方法分析 3 组之间的生存率, 得出生存曲线(图 2). 结果 *CREB5*($P=0.0036$), *EHBPI*($P=0.0029$), *GRB10*($P=0.0396$), *PODXL*($P=0.0007$)和 *STX2*($P=0.0018$)基于基因表达值的变化, 生存率的差异较为明显, P 值均小于 0.05. 而 *KLK6*($P=0.1069$)虽然 P 值大于 0.05, 但低表达患者的中位生存期明显高于高表达患者, 且有文献支持此结果, 说明 *KLK6* 同样与患者的生存期相关. 从这 6 个基因的生存曲线结果可以看出, 多数结果显示这些基因低表达的患者生存率高, 而高表达的患者生存率低, 这表明, 这些基因对大肠癌的发生与发展起着积极的影响, 而且从生存曲线可以看出, 6 个基因对肿瘤的影响程度不一致, 其中 *CREB5* 与 *STX2* 在低表达时患者生存率较高, 但中表达时的生存曲线与高表达时相贴近, 生存率均较低, 说明了这些基因对大肠癌的发生发展影响较大, 对患者生存时间影响较大.

2.4 生存相关关键基因的网络调控分析结果

筛选出的 6 个生存相关基因虽有各自相应功能, 但功能间相互协调, 相互交错. 为了进一步明确 6 个生存相关基因在大肠癌细胞信号网络中的相互作用情况, 使用 STRING 工具对上述 6 个基因的相互作用蛋白相互交错情况进行分析. 为了提高相互作用蛋白的可靠性, 参数中选择基于实验的蛋白连接方法, 由于蛋白间网络调控情况复杂, 蛋白相互作用数量过多会使得研究重点不够突出, 因此以最少的蛋白数量尽可能多的连接上述 6 个蛋白, 构建蛋白相互作用网络, 因此控制网络中蛋白数量为 30 个(图 3). 然后将上述网络中的 30 个蛋白节点使用 DAVID 工具^[7]进行 KEGG 通路富集(表 3), 从结果可以看出, 网络中的蛋白主要参与了大肠癌、黏着斑、ERbB 信号通路, 肿瘤等相关信号的调控. 而通过 GO(gene ontology)分析又可以发现, 大部分网络中的蛋白均定位于

细胞膜上, 而且又与细胞分泌相关, 说明网络中的蛋白与细胞黏附密切相关. 因此, 筛选出的 6 个与大肠癌生存相关关键基因所构成的蛋白相互作用网络不仅与大肠癌相关, 更与大肠癌的转移密切相关, 说明 6 个生存相关基因是通过调控大肠癌转移进而影响大肠癌患者生存时间的变化.

3 讨论

大肠癌的转移决定了患者生存时间的长短, 是否发生转移关系到患者的治疗及预后. 通过筛选与大肠癌患者生存期相关的关键基因, 明确了这些基因表达的差异与患者生存时间的关系, 并通过蛋白相互作用网络调控分析明确了这些基因是通过调控转移相关信号通路而发挥作用, 进一步证实了大肠癌转移与患者生存时间的关系. 从筛选出的 6 个基因

表 2 转录因子结合位点数大于 7 个的基因名称

基因符号	TFBS 数目	基因名称
CREB5	13	cAMP 反应元件结合蛋白 5
LOC401317	13	假定蛋白 LOC401317
ETV1	12	Ets 变异基因 1
MEIS2	12	同源盒蛋白 Meis2
PALLD	12	细胞骨架相关蛋白 Palladin
EHBP1	11	EH 域结合蛋白 1
MITF	11	小眼畸形相关转录因子
TNIF	11	TRAF2 和 NCK 激酶相互作用蛋白
TGFB111	10	转化生长因子 β 1 诱导转录因子 1
CALD1	9	钙调素结合蛋白 1
FERMT2	9	丝裂原诱导蛋白 2
ZNF83	9	锌指蛋白 83
ACTA2	8	平滑肌肌动蛋白 α 2
ATF3	8	激活转录因子 3
FN1	8	纤维连接蛋白 1
GRB10	8	生长因子受体结合蛋白 10
LTBP1	8	组织内转化生长因子 β 结合蛋白 1
OSBPL1A	8	氧固醇结合蛋白样蛋白 1A
PAM	8	肽酰甘氨酸 α 酰胺化单氧酶
PDLIM5	8	PDZ 和 LIM 域蛋白 5
STX2	8	突触融合蛋白 2
ARHGAP5	7	Rho GTP 酶激活蛋白 5
CCDC80	7	卷曲螺旋结构域蛋白 80
DPYSL3	7	二氢嘧啶酶相关蛋白 3
ENTPD4	7	外核苷三磷酸二磷酸水解酶 4
HOPX	7	同源盒蛋白 HOP
KLK6	7	激肽释放酶 6
PII5	7	肽酶抑制因子 15
PODXL	7	足糖萼蛋白样蛋白
UTS2	7	尾加压素 2

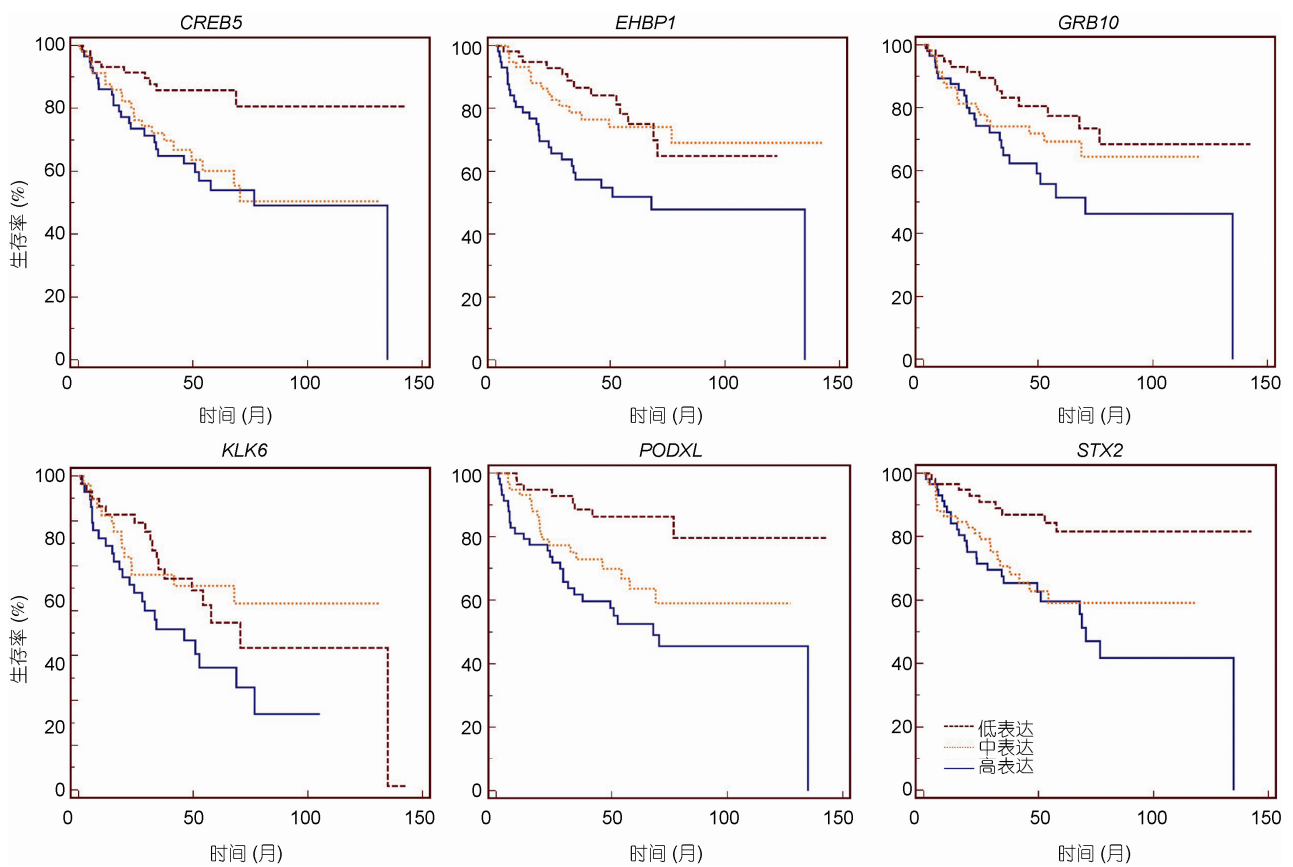


图 2 6 个生存期相关基因生存分析图

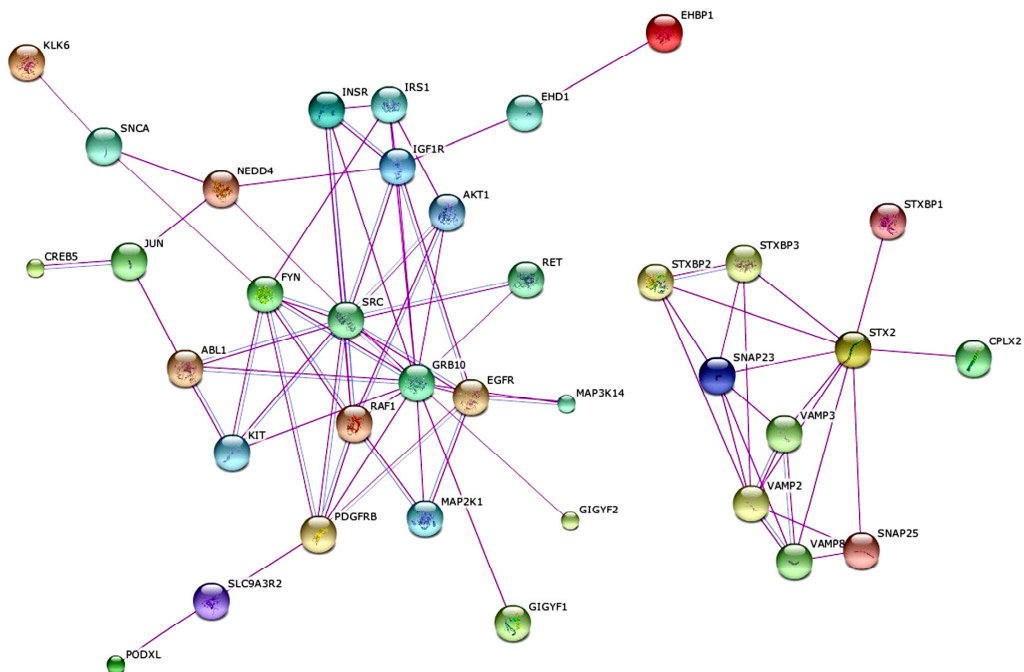


图 3 6 个生存相关基因相互作用网络图

表 3 网络中 30 个相互作用蛋白节点的 KEGG 通路富集结果

分类数据库	类别	基因数目	P 值
KEGG_PATHWAY	hsa05210: 大肠癌	7	1.37×10^{-6}
KEGG_PATHWAY	hsa04510: 黏着斑	9	1.53×10^{-6}
KEGG_PATHWAY	hsa04012: ErbB 信号通路	7	1.69×10^{-6}
KEGG_PATHWAY	hsa05200: 肿瘤相关信号通路	10	6.32×10^{-6}
KEGG_PATHWAY	hsa05214: 神经胶质瘤	6	7.04×10^{-6}
KEGG_PATHWAY	hsa05218: 黑色素瘤	6	1.27×10^{-5}
KEGG_PATHWAY	hsa04130: 膜泡运输中的 SNARE 相互作用	5	2.12×10^{-5}
KEGG_PATHWAY	hsa05215: 前列腺癌	6	3.85×10^{-5}
KEGG_PATHWAY	hsa04660: T 细胞受体信号通路	6	9.77×10^{-5}
KEGG_PATHWAY	hsa04144: 内吞作用	7	1.25×10^{-4}
KEGG_PATHWAY	hsa04722: 神经营养因子信号通路	6	1.88×10^{-4}
KEGG_PATHWAY	hsa04520: 黏着连接	5	3.46×10^{-4}
KEGG_PATHWAY	hsa04540: 间隙连接	5	6.02×10^{-4}
KEGG_PATHWAY	hsa04912: GnRH 信号通路	5	8.67×10^{-4}
KEGG_PATHWAY	hsa04010: MAPK 信号通路	7	9.37×10^{-4}

的生存曲线可以看出,随着基因的表达增高患者的生存率明显下降,说明这些基因的高表达促进了大肠癌的发展.而且 *STX2* 与 *CREB5* 基因由于中表达及高表达时生存曲线重叠,且与低表达时的曲线有着明显的区别,说明这两个基因的中表达即可达到高表达时对患者生存率的影响程度,表明这两个基因的促瘤作用更为强烈.已有报道 *KLK6*^[8-11] 及 *PODXL*^[12-14] 基因在大肠癌组织中高表达,且与患者预后相关,其表达升高患者生存率降低,与本研究分析结果一致,进一步说明了本研究结果的可靠性.生存期与大肠癌患者的治疗及预后息息相关,虽然在前期研究中发现与大肠癌患者生存期相关的基因有

很多,但这些基因并非单独影响着患者生存期,而是通过调控其他相关蛋白进而影响肿瘤细胞的生长和迁移,而其所调控的基因中的许多基因都与肿瘤的转移相关.由于受调控越多的基因,在功能上越重要,因此通过筛选含有转录因子结合位点较多的基因,能够更好的挖掘对患者生存期起着关键作用的基因,且这些基因的表达变化,会较大程度上影响的肿瘤细胞的发展方向,对患者的生存时间及预后有着重要影响.本研究筛选出了 6 个与大肠癌患者生存期密切相关的基因,通过对这些基因表达的检测,有可能对大肠癌患者预后及生存时间进行评估,以期把握最佳的治疗时机.

参考文献

- 1 齐鲁,袁理,吴萍,等. 大肠癌转移相关分子标签的筛选. 现代生物医学进展, 2012, 12: 5225-5229
- 2 Tonon L, Touzet H, Varre J S. TFM-Explorer: mining *cis*-regulatory regions in genomes. Nucleic Acids Res, 2010, 38: W286-W292
- 3 Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res, 2013, 41: D808-D815
- 4 Blanchette M, Bataille A R, Chen X, et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res, 2006, 16: 656-668
- 5 Subramanian A, Tamayo P, Mootha V K, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA, 2005, 102: 15545-15550
- 6 Barrett T, Suzek T O, Troup D B, et al. NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Res, 2005, 33: D562-D566
- 7 Huang D W, Sherman B T, Tan Q, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res, 2007, 35: W169-W175
- 8 Ogawa K, Utsunomiya T, Mimori K, et al. Clinical significance of human kallikrein gene 6 messenger RNA expression in colorectal cancer. Clin Cancer Res, 2005, 11: 2889-2893
- 9 Kim J T, Song E Y, Chung K S, et al. Up-regulation and clinical significance of serine protease kallikrein 6 in colon cancer. Cancer, 2011,

117: 2608–2619

- 10 Petraki C, Dubinski W, Scorilas A, et al. Evaluation and prognostic significance of human tissue kallikrein-related peptidase 6 (KLK6) in colorectal cancer. *Pathol Res Pract*, 2012, 208: 104–108
- 11 Ohlsson L, Lindmark G, Israelsson A, et al. Lymph node tissue kallikrein-related peptidase 6 mRNA: a progression marker for colorectal cancer. *Br J Cancer*, 2012, 107: 150–157
- 12 Barderas R, Mendes M, Torres S, et al. In-depth characterization of the secretome of colorectal cancer metastatic cells identifies key proteins in cell adhesion, migration, and invasion. *Mol Cell Proteomics*, 2013, 12: 1602–1620
- 13 Larsson A, Johansson M E, Wangefjord S, et al. Overexpression of podocalyxin-like protein is an independent factor of poor prognosis in colorectal cancer. *Br J Cancer*, 2011, 105: 666–672
- 14 Larsson A, Fridberg M, Gaber A, et al. Validation of podocalyxin-like protein as a biomarker of poor prognosis in colorectal cancer. *BMC Cancer*, 2012, 12: 282