

# 基于 GEO 数据库对类风湿性关节炎相关基因筛选及生物信息学分析

陈龙梅, 杨振华 (上海市宝山区中西医结合医院检验科, 上海 201900)

**摘要:** **目的** 基于生物信息学筛选类风湿性关节炎 (rheumatoid arthritis, RA) 的差异表达基因, 并分析差异表达基因的生物学功能及其调控通路。**方法** 从 GEO (gene expression omnibus) 数据库检索并下载了基因芯片 GSE94519, 通过 GEO 数据库的在线分析工具 GEO2R 以  $P < 0.05$ ,  $|\log FC| > 1.5$  为条件筛选 RA 的差异基因, 以 DAVID6.8 对筛选出的 RA 差异基因开展 GO 功能注释和 KEGG 信号通路富集分析。通过 STRING 在线分析工具和 Cytoscape 软件挖掘在 RA 生物学过程中发挥至关重要作用的关键基因。**结果** 该研究共发现差异表达基因 278 个, GO 功能在生物学方面主要介导血小板脱颗粒、病毒过程、氧化还原过程和 GTPase 活性正调节的转导过程, 在细胞功能方面主要参与胞外小体、胞浆、薄膜及细胞质的调控, 在分子功能方面主要富集于 GTPase 活性、泛素蛋白连接酶结合、钙黏蛋白结合参与细胞之间的黏附。KEGG 的分析 RA 差异表达的基因结果表明其主要的信号通路是调节氧化磷酸化以及帕金森病, 在蛋白互作网络中筛选出 10 个 Hub 基因分别为 ACTB, RHOA, PPBP, B2M, MT-CYB, PF4, CFL1, MT-ATP6, VCL 和 TPM1。**结论** 利用生物信息学和 R 语言能有效分析 GEO 数据库的原始基因芯片数据, 获得芯片内在的生物学信息; 通过关键差异基因分析不仅能识别目前已知的类风湿性关节炎相关信号通路, 还能发现一些新的通路或生物学过程。

**关键词:** 关节炎; 类风湿; 生物信息学; GEO 数据库

中图分类号: R593.22; R446 文献标识码: A 文章编号: 1671-7414 (2021) 02-049-05

doi:10.3969/j.issn.1671-7414.2021.02.012

## Gene Screening and Bioinformatics Analysis of Rheumatoid Arthritis Based on GEO Database

CHEN Long-mei, YANG Zhen-hua (Department of Clinical Laboratory, the Baoshan District Traditional Chinese and Western Medicine Hospital of Shanghai City, Shanghai 201900, China)

**Abstract:** **Objective** To screen differentially expressed genes in rheumatoid arthritis (RA) based on bioinformatics, and analyze the biological function and regulatory pathway of differentially expressed genes. **Methods** The gene chip GSE94519 was retrieved and downloaded from the gene expression omnibus database (GEO). The differential gene of RA was screened by the online analysis tool GEO2R with  $P < 0.05$  and  $|\log FC| > 1.5$ , and the function annotation of GO and enrichment analysis of KEGG signal pathway were carried out by DAVID6.8. The key genes that play an important role in the biological process of RA were mined by string online analysis tool and Cytoscape software. **Results** In this study, 278 differentially expressed genes were found. Go function mainly mediates the processes of platelet degranulation, virus process, redox process and positive regulation of GTPase activity in biology. It mainly participates in the regulation of extracellular corpuscles, cytoplasm, membrane and cytoplasm in cell function, In the aspect of and cytoplasm in cell and cytoplasm in cell function, In the aspect of molecular function, GTPase activity, ubiquitin protein ligase binding and cadherin binding are mainly involved in cell adhesion. KEGG analysis of RA differentially expressed genes showed that the main signaling pathway was to regulate oxidative phosphorylation and Parkinson's disease. There were ten hub genes were screened out among the protein interaction network: ACTB, RHOA, PPBP, B2M, MT-CYB, PF4, CFL1, MT-ATP6, VCL and TPM1. **Conclusion** Bioinformatics and R language can effectively analyze the original gene chip data of geo database and obtain the biological information inside the chip. Key differential gene analysis can not only identify the known signal pathways of rheumatoid arthritis, but also find some new pathways or biological processes.

**Keywords:** arthritis; rheumatoid; bioinformatics; GEO database

类风湿性关节炎 (rheumatoid arthritis, RA) 主要累及周围关节, 可导致多处周边关节畸形<sup>[1]</sup>。因机体的免疫细胞的异常增殖以及自身凋亡动态的失

衡, 同时激活了异常的氧化应激, 导致多种信号传导通路失衡, 从而引起了 RA 疾病的发生发展<sup>[2]</sup>。随着大数据时代的到来, 迎来了基因芯片数据的大

作者简介: 陈龙梅 (1988-), 女, 硕士研究生, 主管技师, 主要从事临床分子学检验, E-mail: clm2714@126.com。

通讯作者: 杨振华 (1961-), 男, 本科, 主任技师, 主要从事临床分子学检验, E-mail: shbsyzh@126.com。

量发展<sup>[3]</sup>,各种疾病包括RA的基因表达谱也得到了学者的重视。长链非编码RNA(long non-coding RNA, lncRNA)是一类非编码但在调控蛋白合成过程中有着不可或缺功能的RNA,它主要通过特异性地结合其他基因或蛋白质从而起到重要的生物学作用<sup>[4]</sup>。夏燕等人<sup>[5]</sup>注意到有多条lncRNA能够影响RA患者,本研究拟通过从GEO(gene expression omnibus)数据库下载并整理有关RA的生物信息分析的lncRNA基因微阵列数据,通过分析差异表达的lncRNA,探讨lncRNA在RA发生发展的作用,从而探索RA新的治疗靶点。

## 1 材料与方法

1.1 芯片数据来源 通过GEO(gene expression omnibus)数据库,以“Rheumatoid arthritis”为关键词搜索目标,经筛选后,采用由Xu D等提交以GPL20115为平台的基因芯片GSE94519,包含由3例RA患者(GSM2477389~GSM2477391)和3例健康人(GSM2477392~GSM2477394)构成的6例样本,均为lncRNA表达谱数据。

1.2 RA的差异表达基因分析 应用GEO数据库的在线分析工具GEO2R进行分析,GEO2R采用了R语言中GEOquery以及limma程序包,再以 $P < 0.05$ , $|\log FC| > 1.5$ 为筛选条件,挑选RA的差异表达基因。

1.3 RA差异基因的功能分析 采用DAVID在线分析工具(<https://david.ncifcrf.gov/>)对差异基因开展GO功能注释和KEGG信号通路富集分析。GO功能主要包括三个方面,分别是生物学过程(biological process, BP),细胞定位(cellular component, CC)和分子功能(molecular function, MF)。

1.4 RA的差异表达基因所调控蛋白互作网络与关键基因分析 通过蛋白质相互作用数据库STRING11.0(<https://string-db.org/>)探索RA差异基因间编码的蛋白质间的关系,并建立它们间蛋白质互作网络(protein protein interaction network, PPI)。并将PPI结果在Cytoscape软件中打开编辑,以cytoHubba模块计算PPI网络节点的前10名蛋白网络中有较高连接度的hub基因。

## 2 结果

2.1 RA差异基因 根据RA差异基因的筛选条件,在GSE94519基因芯片中共有差异表达基因278个,其中上调54个,下调224个,其火山图见图1。按照差异倍数logFC大小排序,前三个上调基因为微管相互作用和运输1(MITD1),乙酰辅酶A羧化酶 $\alpha$ (ACACA),损伤特异性DNA结合蛋白1(DDBI);前三个下调基因为神经肽原(NPTN),原肌球蛋白 $\alpha$ (TPM1),凝血因子XIII, A1多肽

(F13A1)。

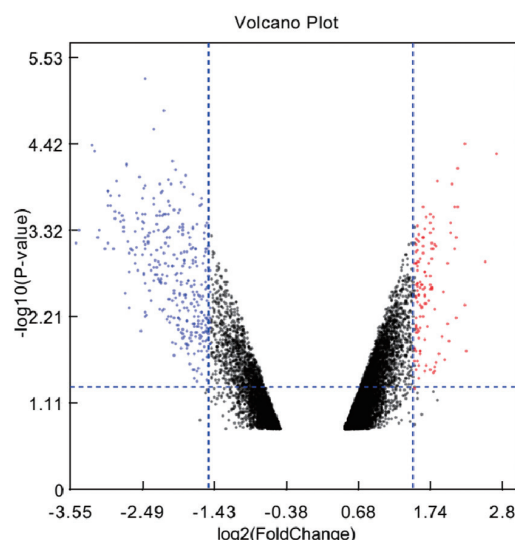


图1 RA差异基因火山图

2.2 RA差异表达基因的GO功能和调控通路分析 RA差异基因在DAVID6.8中的GO功能和KEGG通路分析见表1。GO功能在BP中主要过程为血小板脱颗粒、病毒过程、氧化还原过程、GTPase活性正调节,在CC主要存在于胞外小体、胞浆、薄膜及细胞质,在MF主要发挥GTPase活性、泛素蛋白连接酶结合、钙黏蛋白结合参与细胞间黏附和GTP连接功能。调控的信号主要为调节氧化磷酸化以及帕金森病的信号通路。

2.3 RA差异表达基因所调控蛋白互作网络与关键基因分析 除去孤立无关系的蛋白节点,以RA差异表达基因构建蛋白质相互作用网络,互作关系的蛋白质共251个,构成495条复杂的PPI网络,将网络在Cytoscape软件中进行可视化,见图2。通过cytohubba插件确认Hub基因,结果显示Hub基因分别为ACTB, RHOA, PPBP, B2M, MT-CYB, PF4, CFL1, MT-ATP6, VCL和TPM1,数据显示它们之间可能存在较强的相互作用关系。

## 3 讨论

RA由于其高致残率而引起了社会的广泛关注,但其致病机制尚不明确<sup>[6]</sup>。RA严重影响着患者的生活质量,所以进一步寻找有助于RA早期诊断和治疗的分子靶点至关重要。在本研究中,我们使用了GEO2R在线工具分析了GEO中的芯片数据GSE94519共包含了3例RA患者和3例健康人构成的6例样本。根据筛选条件得到了278个差异表达基因,通过DAVID6.8对差异基因进行GO功能注释和KEGG信号通路分析发现差异基因主要参与了血小板脱颗粒、病毒过程、氧化还原过程、GTPase活性正调节的转导过程和调控氧化磷酸化

和帕金森病等通路。

表 1 RA 差异基因的 GO 功能和调控信号通路的前 20 个基因			
GO 类型	生物学功能	基因数量 (个)	基因名称 (只列举 20 个)
生物学过程 (BP)	血小板脱颗粒	13	CD9, TLN1, PPBP, F13A1, CLU, CALM3, ABCC4, PF4, TMSB4X, ITGB3, SRGN, CTSW, VCL
	病毒过程	13	ATP6V0C, TLN1, VRK2, DYNLL1, DDB1, EIF4A1, HLA-A, RHOA, HLA-C, HLA-B, ZYX, CCNA2, RBX1
	氧化还原过程	15	CYB5R3, ND1, PGD, PTGS1, SNCA, GMPR, MOXD1, FTH1, FAR2, MARC2, PRDX6, GPX4, ABCC4, SMOX, SH3BGRL3
	GTPase 活性正调节	13	ALS2, RGS10, NCK2, RSU1, CALM3, RGS18, GNAS, HERC2, SH3BGRL3, CCL5, TRIP10, TRAPPC1, DENND1B
细胞组分 (CC)	胞外小体	68	CYB5R3, SLC36A2, TLN1, S100A8, PGD, PTGS1, S100A9, MITD1, VCL, B2M, GPX1, GSTM2, DYNLL1, GPX4, RHOA, TUBB1, ATP6, RSU1, DDB1, HLA-A
	胞浆	64	ALS2, SAT1, TLN1, S100A8, SNCA, PGD, S100A9, FOXO3, VCL, GPX1, GSTM2, AP1S2, PRMT2, PBXIP1, DYNLL1, MAP1LC3B, GPX4, RHOA, MGLL, SMOX
	薄膜	46	ALS2, CYB5R3, ATP6V0E1, ABCD1, SNCA, MMD, PPM1A, CYTB, CTSA, FOXO3, B2M, CD9, FUBP3, ATP2B4, STT3A, MCTP2, DYNLL1, RC3H2, MGLL, PCSK6
	细胞质	80	MTRNR2L8, CYB5R3, SLC36A2, TLN1, CDC14B, MTRNR2L2, SNCA, PTGS1, MTRNR2L1, APOBEC3H, FOXO3, MXI1, B2M, GPX1, MAX, GSTM2, FUBP3, MCTP2, PRMT2, DYNLL1
分子功能 (MF)	GTPase 活性	10	GTPBP2, DNM3, RAB31, GNG10, RHOA, RAB11B, GNG11, GNAS, TUBG1, TUBB1
	泛素蛋白连接酶结合	10	ATP6V0C, GABARAPL2, YWHAZ, MAP1LC3B, UBE2K, PRDX6, CLU, HERC2, RBX1, VCL
	钙黏蛋白结合参与细胞-细胞粘附	10	TLN1, YWHAZ, PRDX6, HIST1H3A, HIST1H3B, RAB11B, TAGLN2, CTNNA1, HIST1H3H, VCL
	GTP 连接	10	GTPBP2, DNM3, RAB31, GUCY1A2, RHOA, RAB11B, ARL9, GNAS, TUBG1, TUBB1
信号通路 (KEGG)	氧化磷酸化	15	ATP5E, ND1, ATP6V0E1, ND4, ND5, ND2, ND3, CYTB, ATP6V0C, ND4L, COX2, COX1, ND6, ATP8, ATP6
	帕金森病	14	ATP5E, ND1, ND4, ND5, ND2, ND3, SNCA, CYTB, ND4L, COX2, COX1, ND6, ATP8, ATP6

运用蛋白互作数据库 STRING11.0 以及 Cytoscape 软件分析差异基因,发现 Hub 基因分别为 ACTB, RHOA, PPBP, B2M, MT-CYB, PF4, CFL1, MT-ATP6, VCL, TPM1。 $\beta$ -肌动蛋白 ( $\beta$ -actin, ACTB) 是一种细胞骨架肌动蛋白,其蛋白表达与血小板聚集、凝血、纤维蛋白凝块形成相关<sup>[7]</sup>。RHOA 蛋白是小 G 蛋白超家族的亚家族成员之一,具有 GTP 酶活性,其相关的信号通路在 RA 患者关节成纤维样滑膜细胞 (FLS) 迁移、侵袭、增殖的调控作用较多地被研究<sup>[8]</sup>。原血小板碱性蛋白 (Pro-Platelet basic protein, PPBP) 又名趋化因子 7 (C-X-C motif chemokine ligand 7, CXCL7) 或中性粒细胞活化肽 2 (neutrophil-activating peptide-2, NAP-2), 为 CXC 趋化因子家族的血小板衍生生长因子,能够促进淋巴细胞和内皮细胞激活,

诱导炎症应答,且 PPBP 在 RA 滑膜液中表达出现下调<sup>[9]</sup>。B2M 基因在人体中编码  $\beta$ 2 微球蛋白,RA 患者的血  $\beta$ 2 微球蛋白高于正常人群,这可能与 RA 的炎症途径有关<sup>[10]</sup>。血小板第 4 因子 (platelet factor 4, PF4) 是由血小板  $\alpha$  颗粒合成的一种特异蛋白质,可促进滑膜细胞的增殖与局部抗体的产生等作用,参与 RA 的免疫应答和趋化因子信号通路<sup>[11]</sup>,通过与其它细胞因子相互作用介导 RA 发生发展中骨及软骨的破坏。MT-ATP6 是线粒体 ATP 合酶亚基 6 基因,它是线粒体能量合成过程中的一个关键酶,根据功能有学者预测认为 MT-ATP6 具有潜在的 RA 致病性<sup>[12]</sup>,筛选出的 Hub 基因中 MT-CYB, CFL1, VCL 和 TPM1 目前尚未找到相关的研究,有可能是 RA 潜在的早期诊断的基因靶点,为后续的研究打下基础。



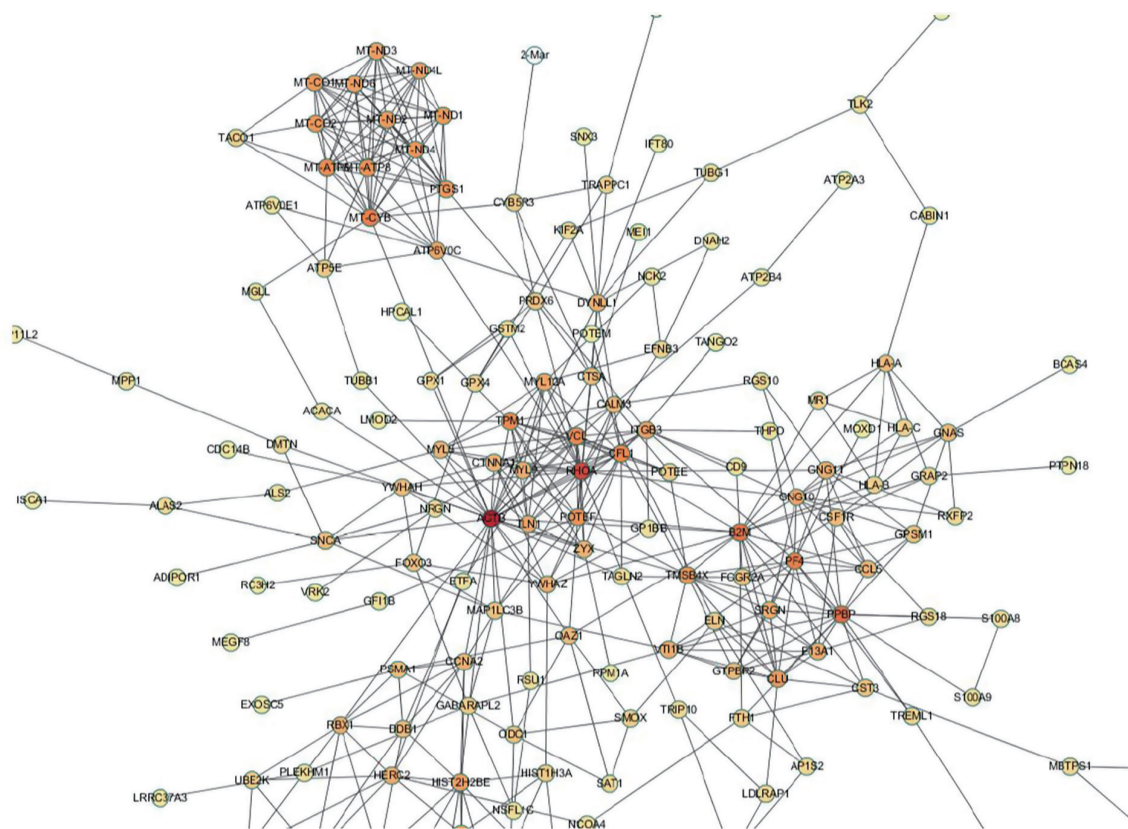


图2 RA 差异基因的 495 条蛋白网络图

本研究是通过探索 GEO 数据库中的芯片 GSE94519 数据, 从而得到 RA 的 Hub 基因, 从而进一步研究 RA 的发病机制、功能和通路分析为 RA 的发生发展寻找新的作用靶点提供了新的思路。

#### 参考文献:

- [1] ZHANG Tianping, ZHANG Qin, WU Jun, et al. The expression levels of long noncoding RNAs lnc0640 and lnc5150 and its gene single-nucleotide polymorphisms in rheumatoid arthritis patients[J]. Journal of Cellular Biochemistry, 2018, 119(12): 10095-10106.
  - [2] 周莹, 王子铭, 虞伟, 等. 类风湿关节炎免疫学发病机制研究的最新进展[J]. 现代检验医学杂志, 2019, 34(1): 157-160.
  - [3] 谢小娟, 朱娜, 潘晶晶, 等. miRNA-148a 在膀胱癌组织中的表达及生物信息学分析[J]. 现代检验医学杂志, 2015, 30(4): 6-9, 13.
  - [4] 吴良银, 李文丽, 刘俊. 肝细胞癌患者生存预后相关长链非编码 RNA(lncRNA) 的生物信息学分析[J]. 现代检验医学杂志, 2019, 34(4): 18-21.
  - [5] 夏燕, 冯佳, 陈安平, 等. 类风湿关节炎外周血 lncRNA 差异表达研究[J]. 中国免疫学杂志, 2016, 32(1): 9-12, 18.
  - [6] 刘小莉, 张静, 张红梅. 类风湿关节炎患者血浆 FDP 和 DD 水平与疾病活动性的相关性研究[J]. 现代检验医学杂志, 2020, 35(2): 60-64.
  - [7] 王新贤, 殷海波, 姜泉, 等. 基于 iTRAQ 蛋白质组学技术筛选类风湿关节炎湿热痹阻证血清标志物[J]. 中国中西医结合杂志, 2019, 39(10): 1209-1213.
- analysis of long-chain non-coding RNA related to survival and prognosis in patients with hepatocellular carcinoma [J]. Journal of Modern Laboratory Medicine, 2019, 34(4): 18-21.
- XIA Yan, FENG Jia, CHEN Anping, et al. Study on expression of long non-coding RNA in rheumatoid arthritis [J]. Chinese Journal of immunology, 2016, 32(1): 9-12, 18.
- LIU Xiaoli, ZHANG Jing, ZHANG Hongmei. Correlation between plasma FDP, DD levels and disease activity in patients with rheumatoid arthritis [J]. Journal of Modern Laboratory Medicine, 2020, 35(2): 60-64.
- WANG Xinxian, YIN Haibo, JIANG Quan, et al. Screening study for serum markers in rheumatoid arthritis patients with stagnant dampness-heat syndrome based on iTRAQ proteomic technology [J]. Chinese Journal of Integrated Traditional and Western Medicine, 2019, 39(10): 1209-1213.