DOI: 10.14188/j.1671-8852.2024.0790
文章编号 1671-8852(2025)05-0637-07

基于生物信息学分析S100A9在真性红细胞增多症并发动脉血栓中的预测作用

刘 禹 刘晓燕 吴三云 熊 蓓 刘 莉 喻亚兰 左学兰
武汉大学中南医院血液科 湖北 武汉 430071

摘要 目的：借助生物信息学方法筛选出真性红细胞增多症并发动脉血栓的核心基因与信号通路，并探讨其可能的机制和临床作用。方法：下载GEO数据库中真性红细胞增多症和急性冠状动脉综合征的基因表达数据集，利用GEO2R筛选出共表达差异基因。借助微生信数据库针对共表达差异基因开展GO及KEGG富集分析。蛋白质-蛋白质相互作用网络通过STRING数据库建立，Cytoscape可视化和挑选。诊断意义通过评估表达水平和绘制ROC曲线来确定。最后通过GeneMINIA和GSEA工具挖掘S100A9共表达基因和富集通路。结果：确定了61个共表达差异基因，从中筛选出8个核心基因进行评估，评估表明S100A9在真性红细胞增多症并发动脉血栓中高表达且有诊断意义，与S100A8和Toll样受体-4密切相关。结论：S100A9在真性红细胞增多症并发动脉血栓中高表达，可能以S100A8/A9的形式通过调节白细胞和/或Toll样受体信号通路发挥促炎活性来促进真性红细胞增多症患者动脉血栓形成，成为预测真性红细胞增多症并发动脉血栓的生物标志物。

关键词 真性红细胞增多症；动脉血栓形成；急性冠状动脉综合征；S100A9；生物信息学
中图分类号 R555.1
文献标识码 A
开放科学(资源服务)标识码(OSID)：

Predictive role of S100A9 in polycythemia vera complicated with arterial thrombosis based on bioinformatics

LIU Yu, LIU Xiaoyan, WU Sanyun, XIONG Bei, LIU Li, YU Yalan, ZUO Xuelan
Dept. of Hematology, Zhongnan Hospital of Wuhan University, Wuhan 430071, Hubei, China

Abstract

Objective: To screen the key genes and signaling pathways of polycythemia vera complicated with arterial thrombosis by bioinformatics, and to explore their possible mechanisms and clinical effects. **Methods:** The gene expression datasets of polycythemia vera and acute coronary syndrome were retrieved from the gene Expression Omnibus (GEO) database, and differentially expressed genes with co-expression patterns were identified using GEO2R. GO and KEGG enrichment analysis of co-expressed differentially expressed genes were performed with the help of the Wei Sheng Xin database. The protein-protein interaction (PPI) network was established using the STRING and was subsequently visualized using Cytoscape to screen the key genes. Diagnostic significance was determined by evaluating expression levels and drawing ROC curves. Finally, S100A9 co-expressed genes and enriched signaling pathways were mined by GeneMINIA and GSEA tool. **Results:** Sixty-one co-

课题来源：武汉大学中南医院科技成果转化面上项目(编号：2022CGZH-MS010)
作者简介：刘禹，女，1999—，医学硕士生，主要从事骨髓增殖性肿瘤研究，E-mail: 1951929760@qq.com
通信作者：左学兰，女，1963—，医学博士，主任医师，硕士生导师，主要从事髓系肿瘤的基础与临床研究，E-mail: zuoxuelan@whu.edu.cn

expressed differentially expressed genes were identified, and 8 key genes were screened for evaluation. The evaluation revealed that S100A9 exhibited significantly elevated expression in cases of polycythemia vera complicated with arterial thrombosis, demonstrating diagnostic significance. This elevated expression was found to be closely associated with S100A8 and Toll-like receptor-4. **Conclusion:** S100A9 is highly expressed in polycythemia vera complicated with arterial thrombosis. It may play a pro-inflammatory activity in the form of S100A8/A9 to promote arterial thrombosis in patients with polycythemia vera by regulating white blood cell and/or Toll-like receptor signaling pathway. Therefore, S100A9 holds potential as a biomarker for predicting the risk of arterial thrombosis in polycythemia vera.

Key Words Polycythemia Vera; Arterial Thrombosis; Acute Coronary Syndrome; S100A9; Bioinformatics

真性红细胞增多症(polycythemia vera,PV)归类于费城染色体阴性的骨髓增殖性肿瘤(myeloproliferative neoplasms,MPN),是一种造血干细胞克隆性肿瘤。除PV外,原发性血小板增多症和原发性骨髓纤维化也属于经典型MPN。截至目前,研究已揭示MPN存在一个共同的发病机理,即信号传导通路JAK-STAT的激活,这种激活是由造血干/祖细胞中驱动基因突变引起的,如:JAK2V617、CALR和MPL。信号通路的激活又引起骨髓细胞过度增殖,导致外周血成熟细胞增多,但其发病机制尚未完全明确。血栓栓塞是MPN主要并发症,是MPN最常见的致残和致死原因之一[1-2]。在MPN的三种亚型中,PV患者血栓栓塞发病率最高,约为46%,其中动脉血栓的发生率明显高于静脉血栓,约为后者的2~3倍。急性冠状动脉综合征(acute coronary syndrome,ACS)作为动脉血栓中的严重类型之一,成为了PV患者最为常见的致死原因[3-4],动脉血栓是ACS对PV患者的安全和生活质量构成的重大风险,对社会造成经济和公共卫生负担,尤其值得高度注意。为了提高PV患者的预后和生存率,评估动脉血栓形成对这些个体的影响至关重要。本研究应用生物信息学方法探索PV共病动脉血栓形成相关基因。基于NCBI-Gene Expression Omnibus(GEO)公共数据库,探究PV发生与ACS的共同差异表达基因,以及这些基因所可能关联的信号传导通路。在此基础上,我们构建了蛋白质-蛋白质相互作用(protein-protein interaction, PPI)网络,旨在识别并筛选出对于PV并发动脉血栓具有潜在指示作用的核心基因,从而为该共病的诊治提供新的线索与帮助。

1 材料与方法

1.1 数据的收集及处理 在GEO数据库(https://www.ncbi.nlm.nih.gov/geo)中以“Polycythemia vera”和“Acute Coronary Syndrome”为关键词的搜索基因表达数据集,从中下载4个人类基因表达数据集。PV患者和对照组的基因表达数据由GSE103237[5](来自26名PV患者和15名对照个体的$CD34^+$ T细胞)和GSE26049[6](来自41名PV患者和21名对照个体的全血)提供。动脉血栓患者和对照组的基因表达数据由GSE66360[7](来自49名ACS患者和50名对照个体的内皮细胞)和GSE24519(来自34名ACS个体和4名对照个体的全血)提供。

1.2 差异表达基因的筛选 运用GEO2R对4个数据集的疾病组与对照组进行差异基因分析。在筛选过程中,为确保识别出表达变化较大且具有显著统计学意义的差异基因。满足特定筛选条件的基因才被定义为差异表达基因:差异倍数(Fold Change,FC)的对数值$\log_2$FC绝对值>1,即$|\log_2 FC|>1$,同时调整后的$P<0.05$。使用在线维恩(Venn)图工具(http://sangerbox.com/home.html)将PV和ACS的上调和下调差异基因分别取交集,即可获得PV发生与血栓形成表达一致的差异基因。

1.3 GO功能注释及KEGG信号通路富集分析 将前述的共表达差异基因集导入微生信在线数据库[微生信-在线生物信息学分析、可视化云平台(bioinformatics.com.cn)]进行基因本体(geneontology,GO)功能富集分析和京都基因和基因组百科全书数据库(Kyoto encylopedia of genes and genomes,KEGG)信号通路富集分析。为筛选出具有显著统计学差异的主要通路,设定$P<0.05$为筛选标准。

1.4　PPI 网络构建及核心基因筛选　将所选差异共表达基因输入到 STRING 数据库（https://cn.string-db.org/），设定阈值为综合评分＞0.4 以创建 PPI 网络，综合评分反映了两个基因之间相互作用的可信度，评分越高，相互作用越可靠。通过设定这个阈值，能够过滤掉那些相互作用较弱或不显著的连接，从而构建一个更为准确和有意义的 PPI 网络。随后将从 STRING 中获取的数据输入 Cytoscape 3.10.2，可视化 PPI 网络。同时，借助 Cytoscape 的插件 CytoHubba 来提供一系列算法用于评估网络中节点的中心性以此筛选出核心基因。

1.5　核心基因的表达和受试者工作特征的评估　首先在 GSE103237 和 GSE66360 数据集中评估核心基因的表达，然后在 GSE26049 和 GSE24519 数据集中确认。采用 Wilcoxon 检验对数据集中核心基因的表达进行比较，$P<0.05$ 表明差异有统计学意义。使用 SPSS 25.0 制作受试者工作特征（ROC）曲线，分别评价核心基因对 PV 和 ACS 诊断的诊断价值。曲线下面积（*AUC*）＞0.7 表明存在显著差异。

1.6　基因共表达网络　使用 GeneMINIA（https://genemania.org/）的共表达模块，绘制 S100A9 基因共表达网络，并对共表达基因的主要功能进行注释。

1.7　GSEA 分析　使用 GSEA 软件（version 3.0）根据 S100A9 的表达水平将样本分成高表达组（≥50%）和低表达组（＜50%），并从 Molecular Signatures Database（http://www.gsea-msigdb.org/gsea/downloads.jsp）下载了 c2.cp.kegg.v7.4.symbols.gmt 子集合，用以评估相关途径和分子机制，基于基因表达谱和表型分组，设定最小基因集为 5，最大基因集为 5 000，1 000 次重抽样，同时满足归一化富集评分（normalized enrichment score，NES）＞1、$P<0.05$ 和错误发现率（false discovery rate，FDR）＜0.25 时具有统计学意义。

2　结果

2.1　筛选 PV、ACS 的差异基因　在火山图中，符合筛选条件的上调和下调差异基因分别用红色和蓝色表示。在 GSE103237 PV 疾病组与对照组中得到上调及下调差异表达基因各有 453 个和 264 个（图 1A），在 GSE66360 ACS 疾病组与对照组中得到表达上调及下调基因各有 776 个和 291 个（图 1B）。将所有的基因名称转化为统一的 Genesymbol 后，GSE103237 中得到表达上调及下调的 Genesymbol 各有 300 个和 187 个；GSE66360 中得到表达上调及下调的 Genesymbol 各有 488 个和 202 个。用 Sangerbox 将两个数据集的上调和下调差异基因的 Genesymbol 分别用 Venn 图取交集，得到 60 个上调的和 1 个下调的 PV 与 ACS 共同表达差异基因（图 1C、1D）。

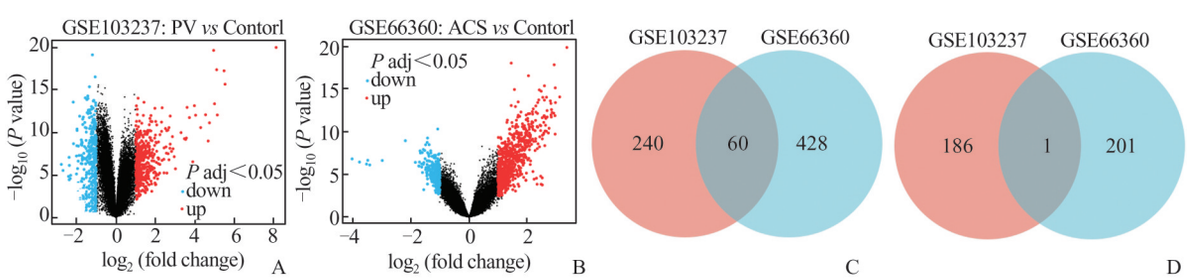


图 1　差异基因火山图和 Venn 图

A. 数据集 GSE103237 PV 疾病组与对照组的火山图；B. 数据集 GSE66360 ACS 疾病组与对照组的火山图；C. 两组上调的差异基因 Venn 图；D. 两组下调的差异基因 Venn 图

2.2　GO 及 KEGG 信号通路富集分析　为了深入分析 PV 发生和 ACS 有关的 61 个共表达差异基因，将其导入至微生物信息学数据库中，进而执行这些差异基因的 GO 和 KEGG 富集分析。GO 富集分析包含“生物过程（biological process，BP）”“细胞组成（cell component，CC）”“分子功能（molecular function，MF）”。差异基因富集的生物过程有中性粒细胞活化，脱颗粒和吞噬作用等（图 2A）；在细胞组成方面，差异基因主要富集于囊泡管腔、胞质囊泡管腔、分泌颗粒管腔、特异性颗粒等颗粒和颗粒管腔（图 2A）；在分子功能方面，差异基因主要参与晚期糖基化终末产物（RAGE）受体结合、模式识别受体活性、Toll 样受体（TLR）结合等模式识别受体活动（图 2A）。KEGG 富集分析提示，差异基因与吞噬体、中性粒细胞胞外陷阱形成等有关（图 2B）。

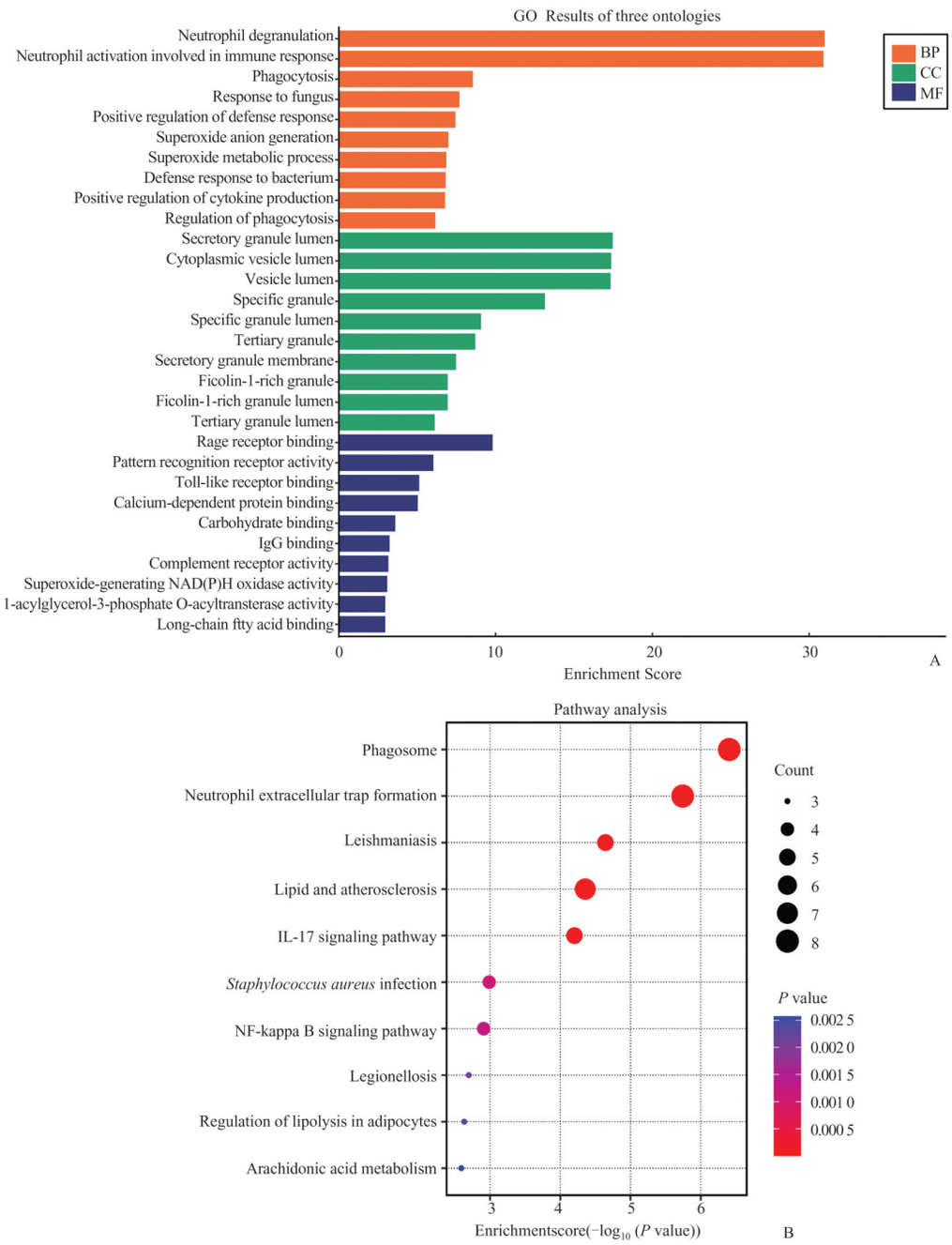


图2 PV与ACS发生共表达差异基因GO功能富集和KEGG信号通路富集结果

A. GO生物过程、细胞组成、分子功能富集结果;B. KEGG分子功能富集结果

2.3 PPI网络构建及核心基因筛选 根据设置的阈值,用STRING数据库分析出以上61个基因之间的PPI网络(图3A),将STRING所得数据输入Cytoscape 3.10.2进行可视化,核心基因运用插件CytoHubba挑选。由于生物网络是异质性的,因此同时使用MCC、DMNC、MNC、Degree和EPC5种拓扑分析算法来预测和探索PPI网络中名次前15的重要核心基因。这5种算法的交集筛选出了8个候选核心基因:S100A9、CYBB、TYROBP、FPR1、NCF2、MNDA、FCN1、S100A8(图3B)。

2.4 确定具有最佳诊断价值的核心基因 我们检查了ACS数据集(GSE66360和GSE24519)和PV数据集(GSE103237和GSE26049)患者组中潜在生物标志物的表达,并将它们与对照组进行了比较(图4A)。结果显示,S100A9、FPR1以及S100A8这三个基因在所有数据集的实验组与对照组之间

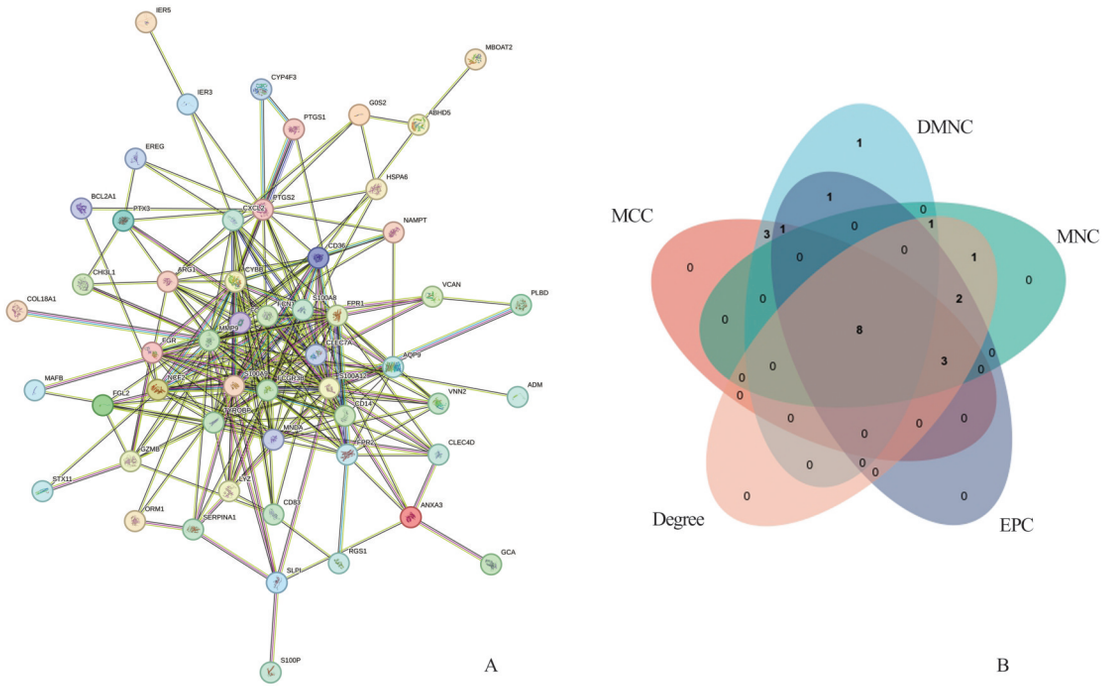


图 3　PPI 网络图及核心基因筛选 Venn 图

A. 共同差异基因的 PPI 网络图；B. MCC、DMNC、MNC、Degree 和 EPC 5 种拓扑分析算法下，PPI 网络中排名前 15 位的重要核心基因 Venn 图

均展现出显著的表达差异（$P<0.05$），这一发现与差异分析的结果保持一致，在 PV 实验组中，这些基因的表达水平呈现上升趋势。在 4 个数据集中绘制了带有 8 个核心基因表达的 ROC 曲线，以评估诊断特征的准确性（图 4B）。在所有数据集中，S100A9 的 AUC 均＞0.7。这些结果表明 S100A9 可能成为诊断 PV 和 ACS 的有前途的标志物。

2.5　基因共表达网络分析　利用 GeneMINIA 探索并构建了 S100A9 的基因共表达网络，并对几个关键功能进行注释（图 5）。结果显示，共挖掘出 20 个共表达基因，且 S100A8 与 S100A9 密切相关，显示与 S100A9 类似的生物学过程，主要涉及模式识别受体信号通路、粒细胞趋化性、炎症反应调节和肿瘤坏死因子调节等过程。

2.6　通过 GSEA 分析 S100A9 相关信号通路　根据 NES 评分，纳入了 GSE26049 和 GSE66360 富集程度较高的 10 条信号通路（图 6A、6B）。两者共同富集的通路为 TLR 信号通路。

3　讨论

PV 的血栓形成是一个重大的公共卫生问题，且造成患者的经济和心身负担，是 PV 患者致残和致死的主要危险因素。在 PV 血栓形成中，动脉血栓最为常见，ACS 是最常见的死亡原因。目前 PV 动脉血栓发生的发病机制尚未明确，可能是由 JAK2V617 突变诱导的白细胞数量和功能异常、慢性炎症和免疫功能障碍所引起的[8]。因此这项研究整合了大量 RNA 数据，应用生物信息学方法来发现 PV 共病动脉血栓的潜在生物标志物，并探讨其中可能的机制和临床作用。

我们的研究从 PV 和 ACS 数据集中确定了 61 个共同表达差异基因，对这些差异基因进行 GO 和 KEGG 分析表明 PV 和 ACS 共病与中性粒细胞和模式识别受体密切相关。中性粒细胞可能通过与模式识别受体结合导致中性粒细胞吞噬、脱颗粒以及中性粒细胞胞外陷阱形成，最终产生免疫紊乱和慢性炎症，从而参与 PV 和 ACS 共病的发生。

将 61 个差异基因导入 STRING 进行蛋白互作网络分析，通过 Cytoscaped 的 CytoHubba 插件，得到 8 个与 PV 和 ACS 共病相关的基因。对 8 个基因的表达和 ROC 曲线进行评估，最终结果表明 S100A9 可能成为诊断 PV 和 ACS 的有前途的标志物。S100A9 是一种主要由粒细胞（如中性粒细胞）表达和分泌的钙锌结合蛋白，在炎症和免疫反应的调节中扮演着至关重要的角色，能够诱导中性粒细胞趋化性、黏附、脱颗粒，增加中性粒细胞的杀菌活性。研究[9-10]表明，S100A9 在 PV 中表达水平增高，且在炎症过程和免疫反应的调节中起着重要作用。在

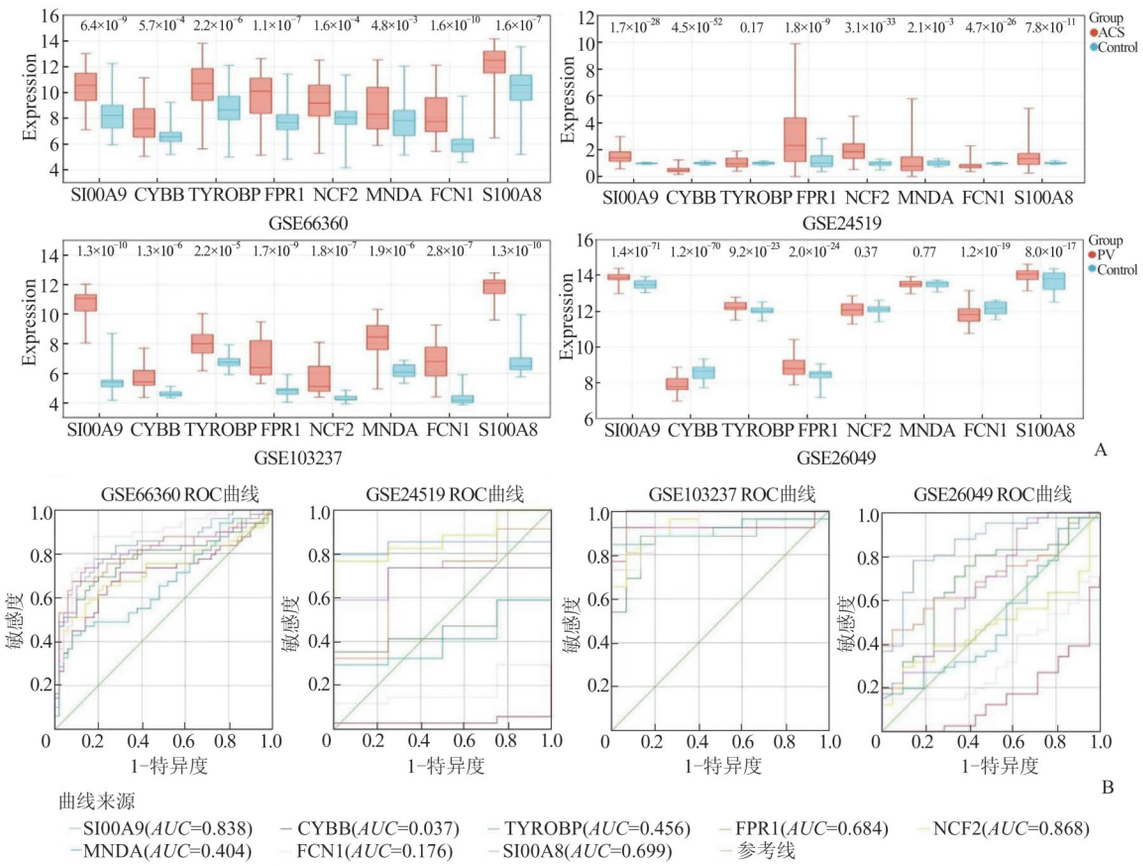


图4 8个核心基因的表达箱式图及ROC曲线

A. 8个核心基因在GSE66360、GSE24519、GSE103237和GSE26049中的表达箱式图；B. 8个核心基因在GSE66360、GSE24519、GSE103237和GSE26049中的ROC曲线

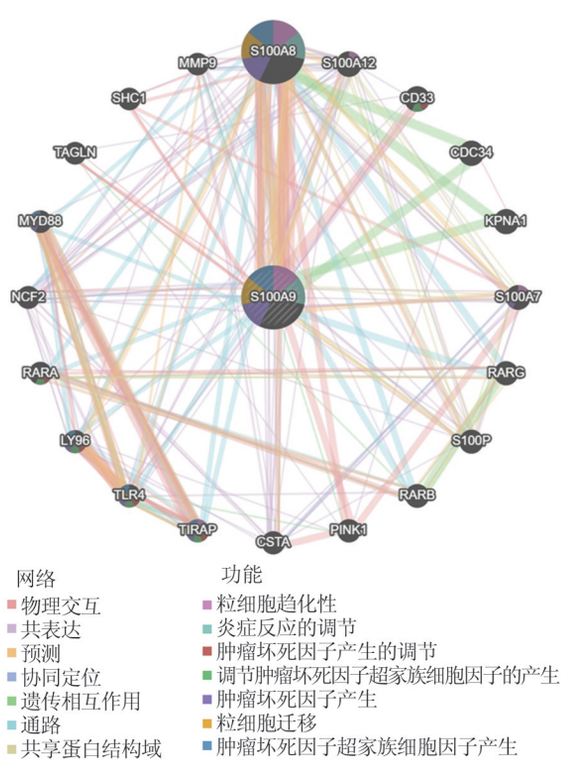


图5 S100A9基因共表达网络

血栓形成的患者中，S100A9表达水平增高，已经有针对S100A9的治疗性疫苗可抑制血栓形成[11]，这与我们的研究结果一致。

通过GeneMINIA在线工具挖掘S100A9的基因共表达网络，结果发现两个重要的相关基因S100A8和Toll样受体4(TLR-4)。S100A8与S100A9密切相关，有类似的生物学过程，主要涉及模式识别受体信号通路、粒细胞趋化性、炎症反应调节和肿瘤坏死因子。S100A8与S100A9都主要由粒细胞(如中性粒细胞)表达和分泌，属于EF-hand钙结合蛋白S100家族，钙卫蛋白(S100A8/A9)是二者的主要发现形式，参与多种功能。S100A8/A9参与激活中性粒细胞NADPH氧化酶和调节性T细胞分化，通过募集白细胞、引导细胞因子产生，并调控白细胞黏附及迁移发挥促炎活性，通过与模式识别受体结合刺激先天免疫细胞，导致促炎级联反应的放大。文献[12-13]表明在骨髓增生性肿瘤中，模式识别受体可以与S100A8/9结合干预骨髓增生性肿瘤中的增殖相关信号。且

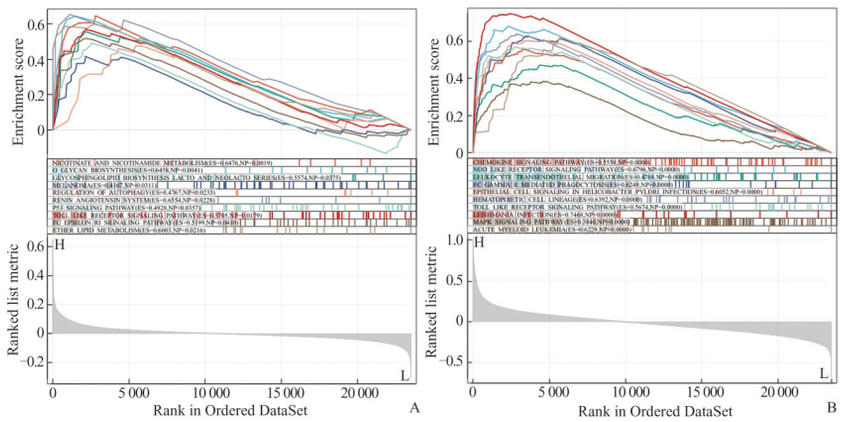


图 6　S100A9 的 GSEA 富集结果

A. S100A9 在 GSE26049 中的 KEGG 通路 GSEA 富集结果；B. S100A9 在 GSE66360 中的 KEGG 通路 GSEA 富集结果

S100A8/A9 与模式识别受体的结合与血栓形成相关。如：S100A8/A9 与库普弗细胞上的 RAGE 受体相互作用导致血小板增多[14]，血小板在血栓形成中起关键作用。在动脉血栓形成中，S100A8/A9 在血管损伤后上调，激活 TLR-4 导致丝裂原活化蛋白激酶及环氧化酶-2 表达升高[15]。

对 GSE26049 和 GSE66360 数据集的基因表达数据进行 GSEA 分析表明 S100A9 可能调控 TLR 信号通路，这与基因共表达结果相呼应。TLR 属于模式识别受体，GO 富集分析表明 PV 和 ACS 共病可能与模式识别受体活性和 TLR 结合有关。以上结果表明 TLR 信号通路可能是 S100A9 调控 PV 并发血栓的关键通路。研究[16]表明在动脉血栓中，TLR-4 可以促进动脉血栓发生，但在 PV 中，TLR-4 对血栓的作用仍不确定。在 MPN 患者中，既往研究[9]表明 TLR2 而不是 TLR4 在 MPN 患者的血栓形成中起重要作用。而 TLR4 单核苷酸多态性与血栓形成风险相关，存在 TLR4-D299G 多态性 MPN 患者表现出对血栓形成保护作用[17]。因此，需要更多的基础和临床研究来证实 TLR-4 在血栓形成中所发挥的作用。

综上所述，S100A9 在预测 PV 并发动脉血栓中可能扮演重要作用，其可能的机制是 S100A9 以 S100A8/A9 的形成的二聚体的形式通过调节白细胞和/或 TLR 信号通路发挥促炎活性从而在共病中发挥作用。但本研究存在一定局限，主要体现在其构建基础与数据特性上可能因样本数量和多重数据分析的复杂性引发潜在的偏倚，且所采用的数据集大多源自西方国家，可能存在种族差异。因此，未来的研究亟须纳入更广泛全面的基因数据集，结合临床患者结果来验证该模型的临床适用性，以全面评估 S100A9 在 PV 并发血栓筛查与防治中的实际应用潜力。

参考文献

[1] TEFFERI A. Primary myelofibrosis: 2023 update on diagnosis, risk-stratification, and management[J]. Am J Hematol, 2023, 98(5): 801-821.

[2] TEFFERI A. Primary myelofibrosis: 2021 update on diagnosis, risk-stratification and management[J]. Am J Hematol, 2021, 96(1): 145-162.

[3] SZUBER N, MUDIREDDY M, NICOLOSI M, et al. 3023 mayo clinic patients with myeloproliferative neoplasms: Risk-stratified comparison of survival and outcomes data among disease subgroups[J]. Mayo Clin Proc, 2019, 94(4): 599-610.

[4] CERQUOZZI S, BARRACO D, LASHO T, et al. Risk factors for arterial versus venous thrombosis in polycythemia vera: A single center experience in 587 patients[J]. Blood Cancer J, 2017, 7(12): 662.

[5] ZINI R, GUGLIELMELLI P, PIETRA D, et al. CALR mutational status identifies different disease subtypes of essential thrombocythemia showing distinct expression profiles[J]. Blood Cancer J, 2017, 7(12): 638.

[6] HOLMSTRÖM M O, RILEY C H, SKOV V, et al. Spontaneous T-cell responses against the immune check point programmed-death-ligand 1 (PD-L1) in patients with chronic myeloproliferative neoplasms correlate with disease stage and clinical response[J]. Oncoimmunology, 2018, 7(6): e1433521.

[7] MUSE E D, KRAMER E R, WANG H Y, et al. A whole blood molecular signature for acute myocardial infarction[J]. Sci Rep, 2017, 7(1): 12268.

[8] SELVAKUMAR B, SEKAR P, RANI SAMSUDIN A. Intestinal macrophages in pathogenesis and treatment of gut leakage: current strategies and future perspectives[J]. J Leukoc Biol, 2024, 115(4): 607-619.

(下转第 654 页)