网络出版时间:2025 - 11 - 13 10:00:02 网络出版地址:https://link.cnki.net/urlid/34.1065.R.20251202.1335.022

机器学习联合生物信息学探究系统性红斑狼疮诊断相关生物标志物

唐 然[1,2],蒋格格[1,2],孟祥文[1,2],蔡 政[1,2],金 莉[3],项 楠[3],张 敏[3],贾晓益[1,2]

[[1]安徽中医药大学药学院,合肥 230012;[2]安徽省活性天然产物重点实验室,合肥 230012;[3]中国科学大学附属第一医院(安徽省立医院)风湿免疫科,合肥 230001]

摘要 **目的** 基于机器学习算法和结构生物学预测,筛选系统性红斑狼疮(SLE)的潜在生物标志物,揭示其作用机制,为SLE诊断和治疗提供新靶点。**方法** 利用随机森林(RF)、极限梯度提升算法(XGBoost)、支持向量机(SVM)、最小绝对收缩和选择算子(LASSO)4种机器学习算法,分析基因表达综合数据库GEO(数据集:GSE121239和GSE11907)中SLE患者基因表达数据,筛选关键标志物。收集SLE患者外周血单个核细胞(PBMCs),采用RT-qPCR法检测差异基因的表达水平。利用GSEA富集分析来确定生物标志物相关通路。应用CIBERSORT免疫浸润分析和蛋白互作网络计算样本免疫细胞浸润丰度。分析单细胞数据在免疫细胞中的基因表达特异性,并结合AlphaFold3(AF3)预测相互作用关系。**结果** 多种算法一起筛选出独特的标记基因HERC5;多个数据集的表达分析显示,与正常组相比,HERC5在SLE中高表达($P<0.05$),RT-qPCR验证了相同的趋势($P=0.006\ 2$)。功能富集分析确定SLE中HERC5促进的主要途径为干扰素受体信号通路($P<0.05$)。免疫浸润分析显示HERC5与免疫细胞密切相关(中性粒细胞:$r=0.39$,$P<0.05$;记忆B细胞:$r=0.33$,$P<0.05$;激活的树突状细胞:$r=0.52$,$P<0.05$)。大多数HERC5相关相互作用蛋白与SLE相关,HERC5及其相关基因的潜在转录因子也与免疫反应显著相关。**结论** HERC5基因是SLE重要的生物标志物,其可能通过干扰素通路促进SLE进展,为SLE诊断和治疗提供新靶点。

关键词 系统性红斑狼疮;机器学习;生物信息学;HERC5;干扰素通路;生物标志物

中图分类号 R 285

文献标志码 A **文章编号** 1000 - 1492(2025)12 - 2368 - 10

doi:10.19405/j.cnki.issn1000 - 1492.2025.12.022

自身免疫性疾病是一类机体对自身抗原产生免疫应答进而引发自身组织损害的疾病,病因复杂,涉及遗传和环境等因素[1]。系统性红斑狼疮(systemic lupus erythematosus,SLE)是一种典型的慢性自身免疫性疾病,该病核心特点是免疫系统过度激活,导致皮肤、肾脏等多脏器功能损伤[2]。然而,SLE的发病机制尚不清楚。目前,SLE患者的临床症状轻重不一,且临床表现具有高度异质性,导致其诊断和治疗极具挑战性。

目前,SLE的诊断主要基于欧洲抗风湿病联盟/美国风湿病学会分类标准,结合抗双链DNA抗体、抗核抗体和抗史密斯抗体等血清标志物及临床表现[3]。尽管一些新型基因标志物如黏病毒耐受蛋白2、Deltex E3泛素连接酶和Dickkopf相关蛋白1等[4-6]显示出潜在的诊断价值,但其临床应用仍受限于敏感性和特异性不足等问题,仍需进一步探索

2025 - 08 - 22 接收

基金项目:国家自然科学基金项目(编号:82074090);安徽省教育厅重点项目(编号:2024AH052061、2024AH040154)

作者简介:唐 然,男,硕士研究生;

张 敏,女,副主任医师,硕士生导师,通信作者,E-mail:doczhangmin@ustc.edu.cn;

贾晓益,女,教授,博士生导师,通信作者,E-mail:jiaxy@ahtcm.edu.cn

vascular space invasion, parametrial involvement, and tumor invasion depth ≥1/2 were identified as significant predictors of PNI. The predictive value was the best in the multivariate model (Area under the curve = 0.80). ***Conclusion*** Perineural invasion is an independent risk factor for poor prognosis of cervical cancer patients, and the occurrence of perineural invasion can be effectively predicted by the constructed multivariate mode.

Key words perineural invasion; cervical cancer; clinicopathological features; independent risk factor; prognosis; adjuvant therapy

Fund program Natural Science Research Project of Anhui Educational Committee (No. KJ2019A0288)

Corresponding author Huang Miaomiao, E-mail:huangmiaomiao@ahmu.edu.cn

更稳定、更具特异性的生物标志物,以提高 SLE 的早期诊断和精准分型。基于此,该研究选择基因表达数据库(gene expression omnibus,GEO)平台,运用生物信息学结合机器学习,鉴定 SLE 关键生物标志物,利用逆转录定量聚合酶链式反应(reverse transcription quantitative polymerase chain reaction,RT-qPCR)实验、通路分析、免疫分析、蛋白互作等研究方法探究关键生物标志物的表达、功能与机制,为 SLE 的临床诊断与基础研究提供理论基础。

1 材料与方法

1.1 数据收集与处理

1.1.1 *数据获取* 从公共基因表达数据库 GEO[7](https://www.ncbi.nlm.nih.gov/geo/)数据平台下载 2 个 SLE 相关数据集,GSE121239 数据集包含 292 例 SLE 患者和 20 例健康个体的外周血单个核细胞(peripheral blood mononuclear cell,PBMCs)样本,GSE11907 数据集包含 110 例 SLE 患者和 12 例健康个体的 PBMCs 样本。这些数据集包括 SLE 患者和健康对照的基因表达信息,用于筛选潜在生物标志物。数据预处理通过 R 软件 R 包 limma 3.58.1 完成,包括去除低表达基因、标准化处理和批次效应校正,以确保分析结果的准确性和可靠性,对于重复出现基因保留均值基因。

1.1.2 *临床样本* 所有患者样本均从中国科学技术大学附属第一医院风湿免疫科获取,共计 22 例患者;所有健康对照样本均从中国科学技术大学附属第一医院体检中心获取,共计 10 例健康对照。收集时间为2023 年 9 月—2024 年 12 月。本研究经中国科学技术大学附属第一医院伦理委员会批准(审批号:2023KY283),所有实验方案均符合《赫尔辛基宣言》。所有参与者在样本采集前均签署了知情同意书。

1.1.3 *主要试剂* Bio RT 高灵敏 cDNA 第一链合成试剂盒(货号:BSB40M1,美国 BioFlux 公司),SYBR Green Realtime PCR Master Mix(货号:A4004D,北京毕特博生物技术有限公司),总 RNA 小量抽提试剂盒(货号:RKB28-02,广州美基生物科技有限公司)。

1.1.4 *主要仪器* 荧光定量 PCR 仪(型号:Mx3000P,美国安捷伦科技有限公司),超微量分光光度计(型号:NanoDrop™ Lite,美国赛默飞世尔科技公司)。

1.1.5 PBMCs 分离与 RT-qPCR 实验 PBMCs 提取:将外周血轻铺在淋巴细胞分离液上,3 000 r/min 离心 25 min,离心后上、中层交界处,以单个核细胞为主的白色雾环即单个核细胞。将其吸附到离心管中,加入磷酸盐缓冲液,混匀,1 500 r/min 离心 10 min,沉淀即为 PBMCs。RNA 的提取及浓度测定:使用 TRIzol 试剂提取细胞中总 RNA。在超微量紫外分光光度计上测得 RNA 浓度。选取酶标仪测定 260 nm 和 280 nm 处的吸光度比值在 1.8 ~ 2.0 符合标准的样本。PCR 引物:HERC5(F):5′-GGCCTTATCCATGTCTGGCAA-3′,HERC5(R):5′-ACCACAAGCGACAAATTCAACTT-3′;GAPDH(F):5′-GGAGCGAGATCCCTCCAAAAT-3′,GAPDH(R):5′-GGCTGTTGTCATACTTCTCATGG-3′。反应条件:在 95 ℃下 10 min 进行预变性,再在 95 ℃下 15 s 和在 95 ℃下 40 s 的 40 个循环进行变性和退火。采用 $2^{-\Delta\Delta C_T}$法分析 HERC5 的相对表达水平。

1.2 机器学习模型构建

为筛选 SLE 的潜在生物标志物,分别使用随机森林(random forest,RF)、极限梯度提升算法(extreme gradient boosting,XGBoost)、支持向量机(support vector machine,SVM)和最小绝对收缩和选择算子(least absolute shrinkage and selection operator,LASSO)等机器学习方法。构建 RF 选择特征得到与 SLE 和对照组显著相关的基因的重要性,提高模型的准确性[8]。XGBoost 是一种高效、可扩展的梯度提升算法,通过迭代优化正则化目标函数构建强预测模型,具有高精度、防过拟合和特征重要性分析能力[9]。SVM 是一种线性分类器,使用基于 SVM 的最大间隔原理训练样本,最后选出需要的特征数,找到最佳变量[10]。LASSO 回归方法可以在拟合广义模型的同时进行变量筛选,进行特征选择和预测特征构建[11]。所有模型均通过 R 语言实现,使用 caret 6.0.94 包随机将输入数据拆分为训练集和测试集,二分类结局指标为 SLE 患者和健康样本。对训练集进行模型拟合并进行 10 折交叉验证。指标评价使用 ROC,数据可视化使用 ggplot2 3.4.4 和机器学习 R 包自带函数。特征基因选取重要性得分排序前 10,筛选结果不足 10 个特征选取全部筛选结果。上述 4 种方法中筛选出的特征基因,通过韦恩图显示重叠的基因,在本研究中进一步分析。

1.3 富集分析

基因富集分析(gene set enrichment analysis,GSEA)(https://www.gsea-msigdb.org/gsea/index.jsp)根据分数对全基因组基因进行排序。使用 R 包“clusterProfiler”进行 Reatome GSEA

富集分析,对表达矩阵进行差异分析,得到差异基因和全基因组基因的富集倍数。错误发现率 < 0.05 被认为是一个显著的富集。此外,采用 Pearson 相关性分析探究最佳关键基因表达水平之间的相关性。

1.4 免疫细胞浸润分析(cell-type identification by estimating relative subsets of RNA transcripts, CIBERSORT) CIBERSORT 方法可将每个样本的基因表达数据与特定的基因集进行比较来估计每个样本的得分。基于 CIBERSORT 方法估计每个样本中 22 种免疫细胞得分,根据样本得分进一步探究 SLE 和正常样本间的免疫浸润差异。结合上述机器学习所得到的枢纽基因,利用基因表达和样本得分的相关性分析探究枢纽基因与免疫细胞之间的关联。

1.5 蛋白互作(protein-protein interaction , PPI)网络构建 将交集靶点基因输入 STRING 数据库(https://cn. string - db. org/),物种限定为"Homo sapiens",设置置信度 > 0.4,同时隐藏离散靶点,得到 PPI 网络。另一种 PPI 网络构建方式使用 GeneMANIA 平台(https://genemania. org/),打开网址 https://genemania. org/,输入 gene list 点进 search 构建调控网络,点击圆圈选项更改网络形式,最后下载结果。

1.6 蛋白结构预测与可视化 利用 AlphaFold3(AF3)对 HERC5 与其候选互作蛋白(ISG15、IRF3、UBE2L6 和 IFIT1)进行复合物结构预测。将蛋白序列(来源于 UniProt 数据库)输入 AF3 模型,并采用默认参数设置。预测结果包括单个残基的可信度评分、蛋白 - 蛋白结合的可信度评分和整体复合体的评分。根据界面预测 TM 分数(inter - chain predicted TM score,ipTM)来评估不同链残基之间的相互作用,用来衡量两个蛋白间界面相互作用准确度,数值越高蛋白结合的稳定性越高,并结合结构可视化工具分析关键结合区域。蛋白结构预测结果通过 PyMOL 软件进行可视化,以直观展示蛋白互作的关键区域和可能的结合模式。

1.7 统计学处理 对于两组之间的连续变量比较,如果它们符合正态分布,则应用 t 检验,非正态分布采用 Mann-Whitney U 检验。方差分析适用于 3 组之间的连续变量。Pearson 的分析用于明确基因表达与免疫细胞分数之间的相关性。ROC 曲线分析用于确定研究中确定的诊断指标的诊断性能。所有统计分析均使用 R 语言(4.3.1 版)进行。所有统计分析均为双侧分析,$P < 0.05$ 为差异有统计学意义。实验结果的统计分析通过 R 语言 ggplot2 包(3.5.1 版)和 GraphPad Prism(10.1.2 版)进行数据可视化。

2 结果

2.1 基于机器学习鉴定 SLE 生物标志物 RF 分析:在 GSE121239 和 GSE11907 数据集中分别使用 RF 算法计算基因的重要性分值,并筛选出(GSE121239 筛选 115 个;GSE11907 筛选 41 个)confirmed 基因作为 SLE 潜在生物标志物,模型 AUC 均为 1.000;XGBoost 分析:通过 XGBoost 算法对 2 个数据集的基因表达数据进行训练与测试,筛选出具有高贡献度的前 15 的枢纽基因,模型 AUC 均为 1.000;SVM 分析:在 SVM 分析中,基于基因表达数据训练分类模型,分别对 2 个数据集的重要基因进行评分,模型 AUC 均为 1.000;LASSO 分析:利用 LASSO 回归模型对 2 个数据集分别进行特征选择,筛选出与 SLE 显著相关的关键基因,GSE121239 和 GSE11907 的 AUC 分别为 1.000 和 0.990。见图 1、2。

基因交集筛选:结合 RF、XGBoost、SVM 和 LASSO 4 种算法的筛选结果,GSE121239 筛选得到 231 个 SLE 潜在生物标志物,GSE11907 共鉴定 159 个潜在标志物。进一步对 2 个数据集潜在标志物取交集,最终鉴出唯一的交集基因 HERC5 作为潜在的 SLE 生物标志物(图 3)。

2.2 HERC5 在 SLE 患者中表达情况 基于 GSE121239 和 GSE11907 数据分析表明,相比于正常样本,HERC5 在 SLE 中显著高表达(图 4A、4B)。再收集 SLE 患者和健康对照外周血 PBMCs,针对筛选出来的差异基因 HERC5 采用 RT-qPCR 实验进行验证,结果显示,与健康对照组相比,HERC5 在 SLE 患者中显著增高(图 4C)。

2.3 HERC5 与 SLE 中浸润免疫细胞的相关性 为进一步探究枢纽基因和免疫细胞的关系,随后对 HERC5 与免疫细胞浸润进行分析。结果显示 HERC5 与 SLE 患者中特定的免疫细胞类型之间存在相关性。HERC5 与中性粒细胞($r = 0.39$, $P < 0.05$),记忆 B 细胞($r = 0.33$, $P < 0.05$),激活的树突状细胞($r = 0.52$, $P < 0.05$)等细胞的浸润呈显著正相关,而与静息的 NK 细胞($r = -0.20$, $P < 0.05$),CD8 T 细胞($r = -0.25$, $P < 0.05$)和初始 B 细胞($r = -0.29$, $P < 0.05$)等细胞的浸润呈显著负相关(图 5)。这表明 HERC5 与免疫细胞浸润的关

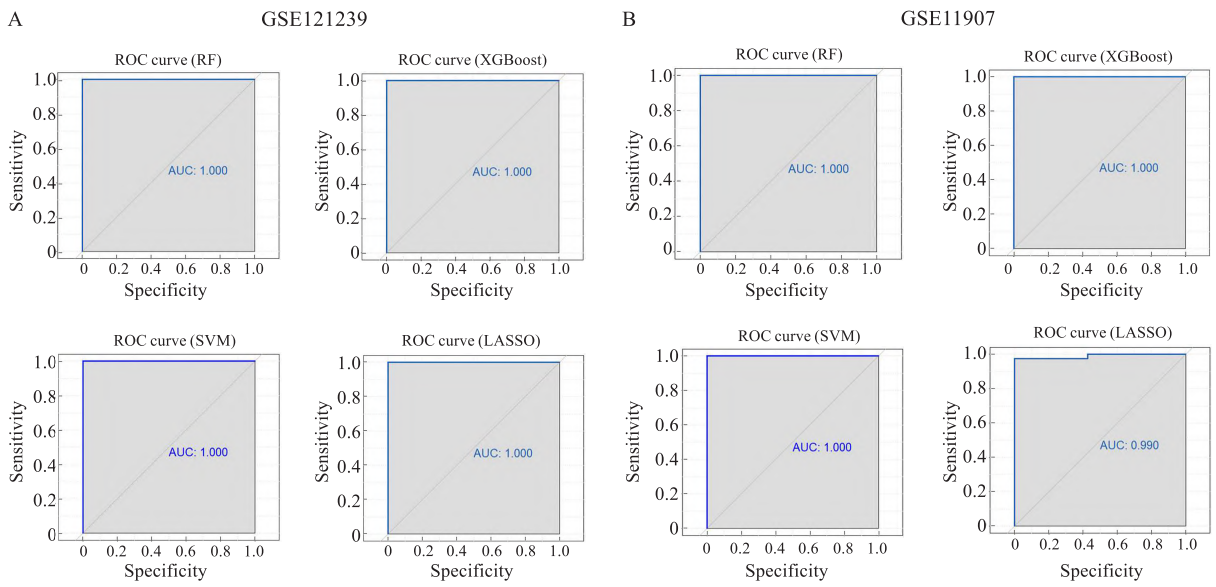

图 1 4 种机器学习算法的验证集 AUC

Fig. 1 Validation set AUC of four machine learning algorithms

A: Validation set AUC of RF, LASSO, SVM and XGBoost (GSE121239); B: Validation set AUC of RF, LASSO, SVM and XGBoost (GSE11907).

系具有显著相关性。

2.4 HERC5 通路分析 为了研究生物标志物 HERC5 在 SLE 中的作用机制,通过 GSEA 的结果发现干扰素信号通路在 SLE 患者中呈现激活趋势(干扰素信号通路归一化富集分数:2.359 833,$P<0.001$;干扰素 Beta 信号通路归一化富集分数:2.808 159,$P<0.001$),提示 HERC5 可能通过调控干扰素信号通路的激活促进 SLE 的发生发展(图 6)。

2.5 HERC5 与蛋白质互作网络 基于 GeneMANIA 平台,构建 PPI 网络。利用 cytoHubba 插件进行分析,HERC5 与多种免疫和炎症相关蛋白(DDX58、PPM1B、FLNB、UBA7、IRF3、IFIT1、ISG15、EIF4G3、EIF4G2、EIF4E3、EIF4E2、HASPIN、UBE2L6、SIRT7、HERC6、TTK、UBE2N、MAP3K14、NOP16 和 AC098582.1)存在相互作用(图 7A)。基于 STRING 平台的 PPI 网络构建显示,HERC5 分别于 IFIT2、DDX58、USP18、ISG15、UBA7、UBE2L6、MX1、IFIT1、OASL 和 IFIT3 存在相互作用(图 7B)。随后将 STRING 和 GeneMANIA 取交集靶点基因,发现 HERC5 与 ISG15、IRF3、UBE2L6 和 IFIT1 有相互作用关系(图 7C、7D)。

2.6 HERC5 与关键蛋白的结构预测 利用 AF3 对 HERC5 与 ISG15、IRF3、UBE2L6 和 IFIT1 的蛋白复合物结构进行预测。预测结果显示,HERC5 与 UBE2L6 的 ipTM 值为 0.83,表明其结合稳定性和可信度较高,提示 HERC5 可能通过参与泛素化途径发挥重要作用;HERC5 与 ISG15 的 ipTM 值为 0.55,结合区域的 pLDDT 值较高,表明该相互作用可能涉及干扰素信号通路的调控。相较之下,HERC5 与 IRF3 和 IFIT1 的 ipTM 值分别为 0.29 和 0.21,可信度较低,可能的结合模式尚不明确。上述预测结果通过 PyMOL 可视化进一步验证了关键结合区域的可靠性(图 8)。提示 HERC5 可能与上述蛋白互作发挥潜在功能。

3 讨论

SLE 是一种发病机制复杂、临床异质性大的自身免疫性疾病,其特征是机体产生攻击自身组织和器官的抗体,导致全身多个器官的炎症反应和组织损伤,其病因目前尚不完全清楚。随着基因芯片技术和高通量技术的发展,利用生物信息学方法挖掘基因芯片数据可以快速有效地筛选差异基因。近年来,它被广泛应用于 SLE 等自身免疫性疾病的致病机制研究,为深入解析其分子病理学基础提供了新的研究思路。

本研究通过对 GSE121239 和 GSE11907 这 2 个基因数据集的生物信息学分析,筛选出 1 个共同差异表达基因 HERC5。HERC5 由 6 个 HERC 蛋白组成的家族,包含1个氨基末端的RCC1样结构域、1

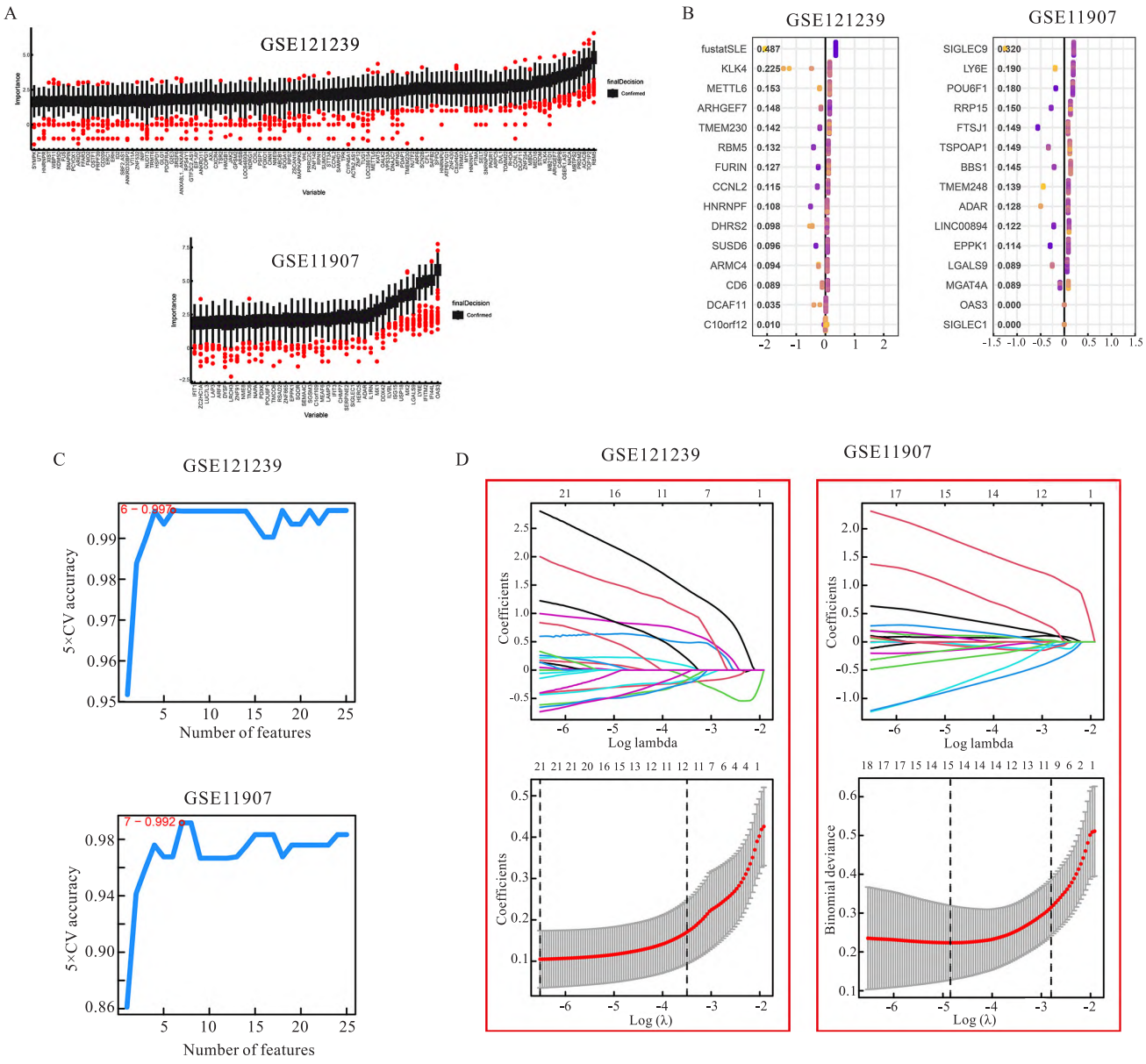

图 2 4 种机器学习算法鉴定 SLE 生物标志物

Fig. 2 Four Machine Learning Algorithms Identify SLE Biomarkers

A: Results of genetic importance analysis of RF algorithm; B: Ranking of gene importance scores for the XGBoost algorithm; C: Results of SVM genetic screening; D: Results of LASSO algorithm gene screening.

个与任何已知蛋白都不具有同源性的间隔区以及 1 个羧基末端的 HECT 结构域。HERC5 蛋白被鉴定为一种抗病毒蛋白,可抑制多种病毒的复制[12-13]。GSEA 结果显示,HERC5 参与干扰素信号通路。Ⅰ型干扰素在先天抗病毒和适应性免疫反应中发挥着重要作用,在微生物感染时迅速发生。HERC5 的表达在体外病毒感染时上调[14]。感染作为 SLE 常见的危险因素,在 SLE 的发生发展过程中发挥着重要的作用。HERC5 在许多细胞类型和组织中普遍表达,包括但不限于效应 T 细胞、中枢记忆 T 细胞、树突状细胞、$CD14^+$单核细胞、单核细胞衍生的巨噬细胞、胚胎干细胞、多能干细胞、造血干细胞[15-17]。在本研究中发现,HERC5 与 SLE 中激活的树突状细胞,记忆 B 细胞,浆细胞和中性粒细胞的免疫浸润呈正相关,而与静息的 NK 细胞,γ-δ 型 T 细胞和初始 B 细胞的免疫浸润呈负相关。Coit et al[18]发现,在肾脏受累的 SLE 患者中,HERC5 存在低甲基化现象。本研究 RT-qPCR 实验也同样发现 HERC5 在

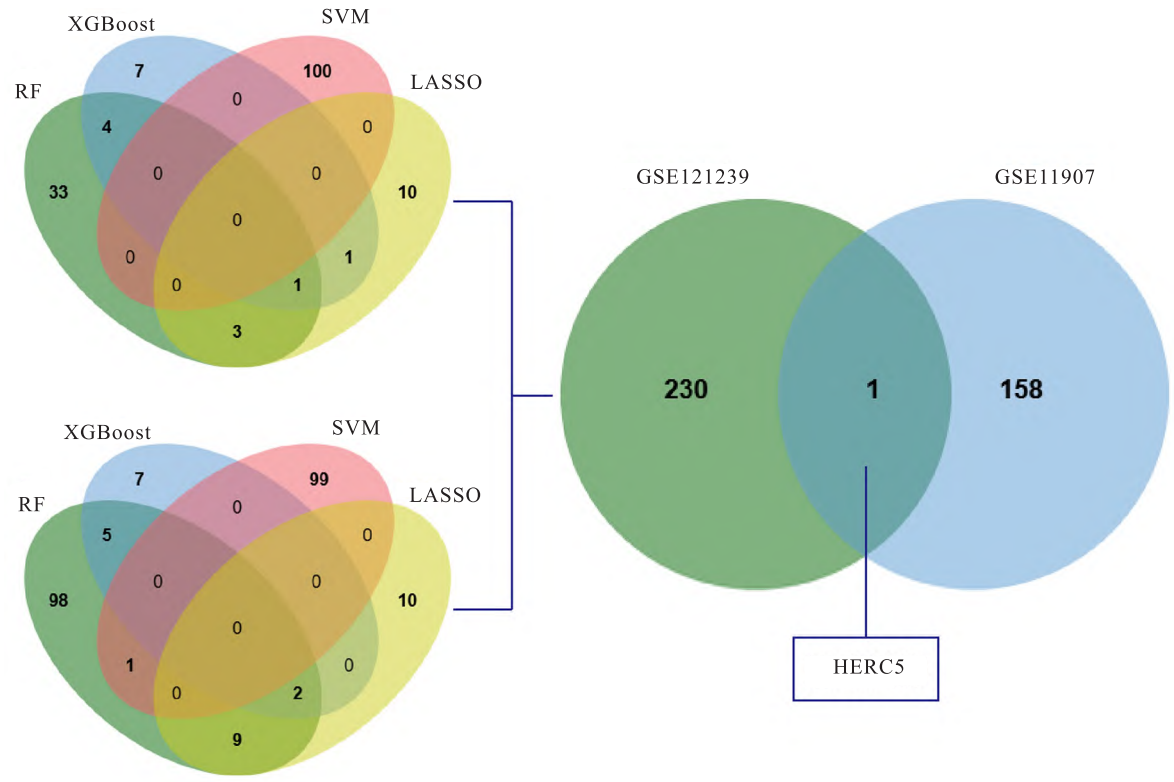

图 3　GSE121239 和 GSE11907 数据集交集基因

Fig. 3　Intersection genes of GSE121239 and GSE11907 datasets

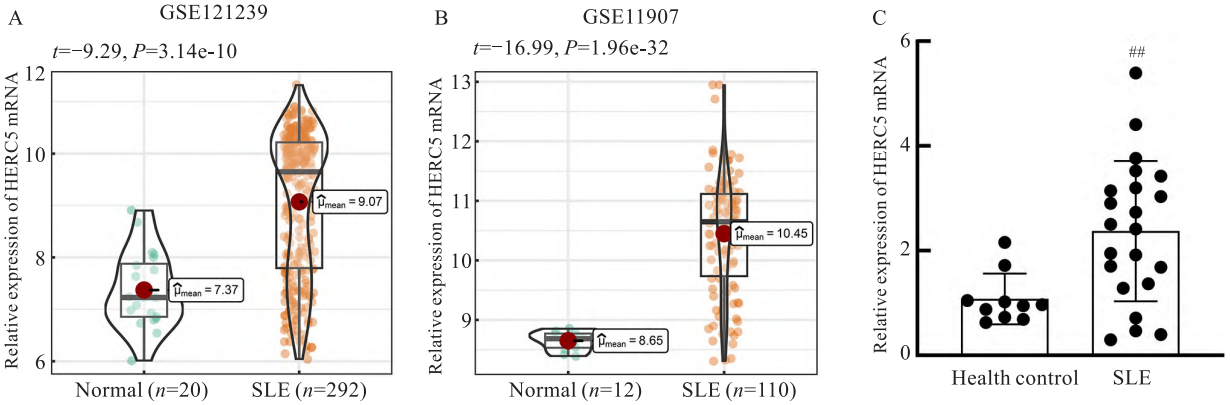

图 4　HERC5 在 SLE 中表达情况

Fig. 4　Expression of HERC5 in SLE

A, B: HERC5 is highly expressed in SLE (GSE121239 and GSE11907); C: mRNA expression of HERC5 in PBMCs; $^{\#\#}P<0.01$ *vs* Health control group.

SLE 患者中高表达。

在 STRING 平台和 GeneMANIA 平台构建 PPI 网络,用 cytoHubba 插件进行分析,得到共同靶点基因 5 个。在上述筛选出的关键基因中,IFIT1 属于 IFIT 家族,是受干扰素诱导产生的一类干扰素诱导基因,在抗病毒和免疫调节中起着重要作用。干扰素最早诱导的 ISG 之一是 ISG15。由 ISG15 基因合成的游离 ISG15 蛋白会在翻译后与细胞蛋白结合,也会被细胞分泌到细胞外环境中。ISG15 蛋白在生理条件下的表达量极低。然而,在多种人类疾病中,包括癌症、神经退行性疾病和炎症性疾病,ISG15 依赖干扰素的表达会异常升高或受损。人类 ISG15 的

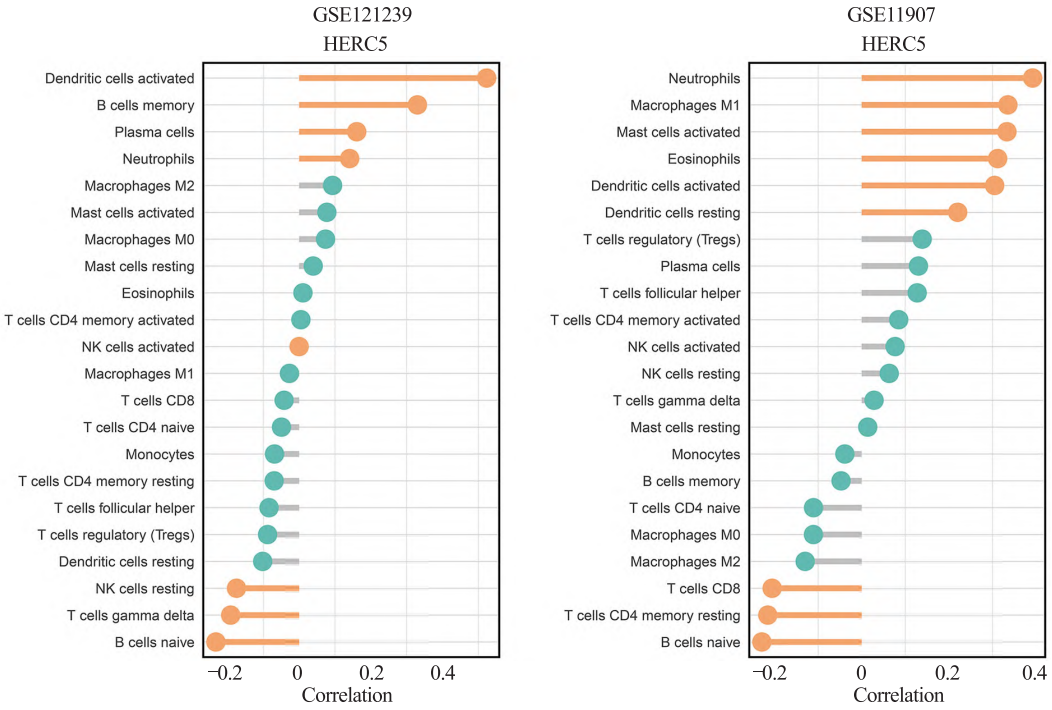

图 5　HERC5 与免疫细胞相关性

Fig. 5　The correlation between HERC5 and immune cell

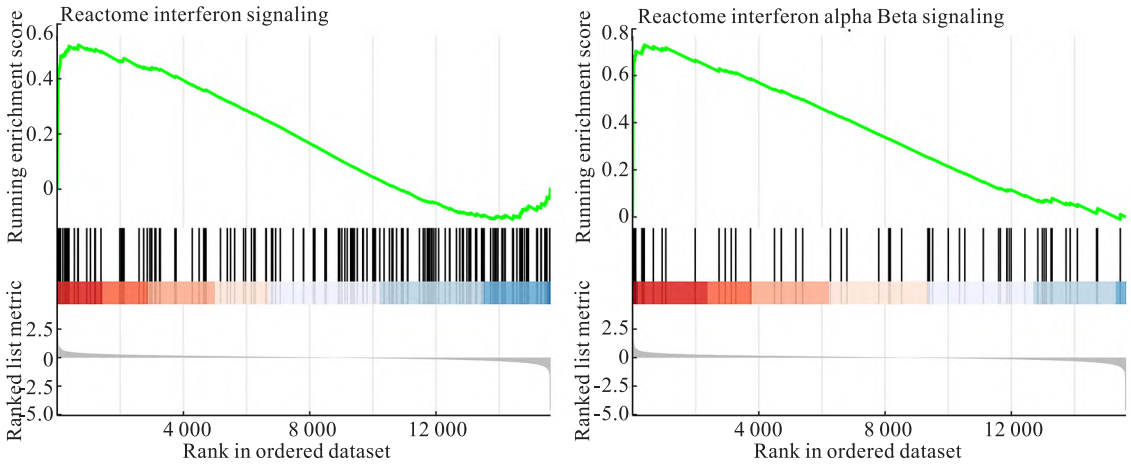

图 6　HERC5 与干扰素信号通路相关性分析

Fig. 6　Correlation analysis of HERC5 and interferon signaling pathway

主要连接酶是 HERC5，可广泛地对蛋白质进行共翻译 ISGylates，ISG15 在人类疾病的病因和发病机制中具有抑制或刺激作用。Ⅰ型干扰素通路被激活后增加了 UBE2L6 的表达，UBE2L6 是一种催化泛素化与其他蛋白质链接的桥梁，是调节蛋白质稳定和功能的重要蛋白。研究[19]发现，UBE2L6 能抑制受结核分枝杆菌感染的巨噬细胞凋亡，miR-146a-5p 可能是 UBE2L6 的靶标。UBE2L6 在全反式维甲酸诱导的急性早幼粒白血病细胞的细胞分化中具有功能性作用，并可能是由其在 ISGylation 中的催化作用介导的[20]。IRF3 是Ⅰ型干扰素产生的主转录因子，IRF3 的转录活性和其他生物功能通过其磷酸化受到精确调控[21-22]。磷酸化的 IRF3 会发生构象变化，随后进入细胞核，与靶基因的启动子结合，从而增强干扰素和干扰素刺激基因的产生[23]。IRF3 与 IRF7 共享 NF-κB 结合基团，可以抑制病毒感染细胞中的炎症基因表达[24]。并且，研究[25]发现 IRF3 是 HERC5 的底物，HERC5 可催化 IRF3 的 ISGylation，

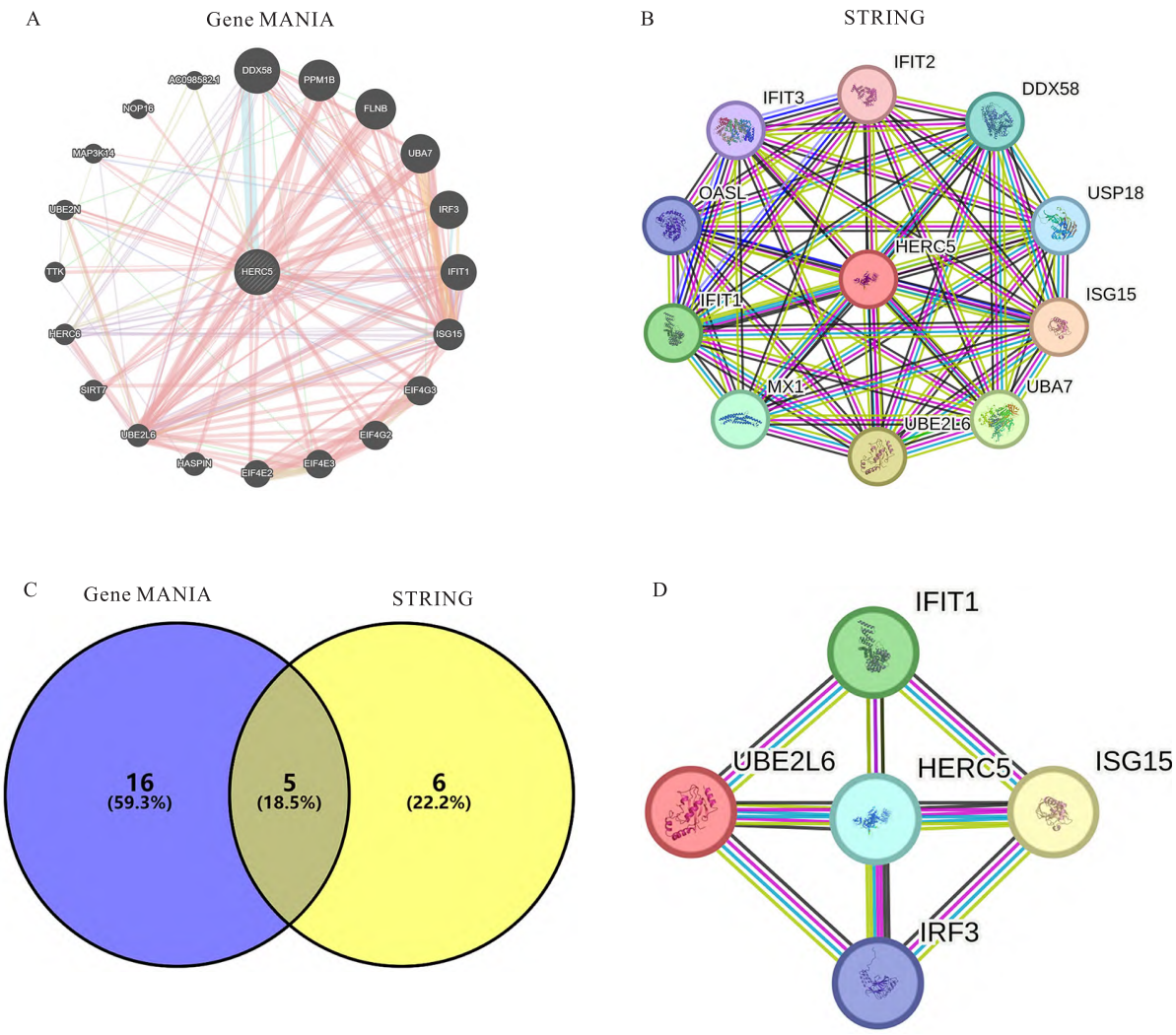

图 7 HERC5 的 PPI 网络图

Fig. 7 PPI network diagram of HERC5

A: Core genes derived from the GeneMANIA platform; B: Core genes derived from the STRING platform; C: GeneMANIA and STRING intersection target Venn diagrams; D: Core target genes associated with HERC5.

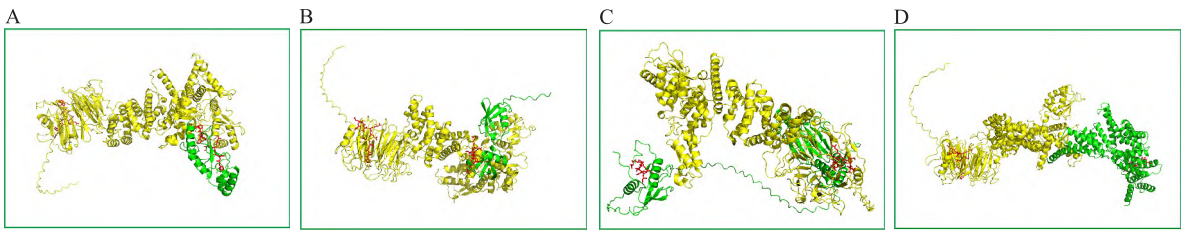

图 8 HERC5 与关键蛋白的可视化

Fig. 8 Visualization of HERC5 with key proteins

A: Visualization of HERC5 with UBE2L6; B: Visualization of HERC5 with ISG15; C: Visualization of HERC5 with IRF3; D: Visualization of HERC5 with IFIT1.

这种修饰抑制了 IRF3 的泛素化和降解,从而增强了先天免疫力。本研究利用 AF3 对 HERC5 与 ISG15、IRF3、UBE2L6 及 IFIT1 的蛋白复合物结构进行预测。结果显示,HERC5 与 UBE2L6 的相互作用具有较高可信度,提示其可能参与泛素化调控;与 ISG15 的中等结合稳定性提示其可能调控 ISGylation 修饰和影响干扰素通路的激活。相比之下,HERC5 与 IRF3 和 IFIT1 的互作预测可信度较低,反映这些相

互作用在生理条件下较弱或需要辅助因子的参与。尽管如此,考虑到 IRF3 是 I 型干扰素通路的主转录因子,而 IFIT1 是典型的干扰素刺激基因,未来研究可通过实验进一步验证 HERC5 是否在特定条件下(如病毒感染或炎症刺激)与这些蛋白发生动态结合。

本研究通过生物信息学分析发现 HERC5 与干扰素通路相关,并证实其在 SLE 患者中高表达,但该研究尚存在一定局限性:该研究尚未直接证实 HERC5 对干扰素通路的调控作用。其次,临床样本数量需要进一步扩大,以深入探究 HERC5 在 SLE 病程中的作用,同时,通过基因干扰、过表达等多种手段,系统评估 HERC5 是否经由 ISGylation 或泛素化途径调节干扰素信号通路。

综上所述,本研究基于通过 RF、XGBoost、SVM 和 LASSO 4 种机器学习算法,筛选出可用于 SLE 诊断的特征基因标志物 HERC5,其可能通过干扰素通路促进 SLE 进展,并通过 RT-qPCR 实验初步验证其参与 SLE 的疾病进程,为 SLE 的早期诊断提供了有价值的研究依据。未来,对 HERC5 的相关机制还需进一步深入研究,以确定其在 SLE 中的具体作用。

参考文献

[1] Miller F W. The increasing prevalence of autoimmunity and autoimmune diseases: an urgent call to action for improved understanding, diagnosis, treatment, and prevention[J]. Curr Opin Immunol, 2023, 80: 102266. doi:10.1016/j.coi.2022.102266.

[2] Accapezzato D, Caccavale R, Paroli M P, et al. Advances in the pathogenesis and treatment of systemic lupus erythematosus[J]. Int J Mol Sci, 2023, 24(7): 6578. doi:10.3390/ijms24076578.

[3] Yu H, Nagafuchi Y, Fujio K. Clinical and immunological biomarkers for systemic lupus erythematosus[J]. Biomolecules, 2021, 11(7): 928. doi:10.3390/biom11070928.

[4] Meng X W, Cheng Z L, Lu Z Y, et al. MX2: identification and systematic mechanistic analysis of a novel immune-related biomarker for systemic lupus erythematosus[J]. Front Immunol, 2022, 13: 978851. doi:10.3389/fimmu.2022.978851.

[5] Leu C M, Hsu T S, Kuo Y P, et al. Deltex1 suppresses T cell function and is a biomarker for diagnosis and disease activity of systemic lupus erythematosus[J]. Rheumatology, 2019, 58(4): 740. doi:10.1093/rheumatology/kez039.

[6] Xue J, Yang J, Yang L, et al. Dickkopf-1 is a biomarker for systemic lupus erythematosus and active lupus nephritis[J]. J Immunol Res, 2017, 2017(1): 6861575. doi: 10.1155/2017/6861575.

[7] Barrett T, Wilhite S E, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update[J]. Nucleic Acids Res, 2012, 41(D1): D991-5. doi:10.1093/nar/gks1193.

[8] Jhansi G, Sujatha K. HRFSVM: identification of fish disease using hybrid Random Forest and Support Vector Machine[J]. Environ Monit Assess, 2023, 195(8): 918. doi:10.1007/s10661-023-11472-7.

[9] Zuo D, Yang L, Jin Y, et al. Machine learning-based models for the prediction of breast cancer recurrence risk[J]. BMC Med Inform Decis Mak, 2023, 23(1): 276. doi:10.1186/s12911-023-02377-z.

[10] Zhu Y X, Huang J Q, Ming Y Y, et al. Screening of key biomarkers of tendinopathy based on bioinformatics and machine learning algorithms[J]. PLoS One, 2021, 16(10): e0259475. doi:10.1371/journal.pone.0259475.

[11] Kang J, Choi Y J, Kim I K, et al. LASSO-based machine learning algorithm for prediction of lymph node metastasis in T1 colorectal cancer[J]. Cancer Res Treat, 2021, 53(3): 773-83. doi:10.4143/crt.2020.974.

[12] Woods M W, Kelly J N, Hattlmann C J, et al. Human HERC5 restricts an early stage of HIV-1 assembly by a mechanism correlating with the ISGylation of Gag[J]. Retrovirology, 2011, 8: 95. doi:10.1186/1742-4690-8-95.

[13] Mathieu N A, Paparisto E, Barr S D, et al. HERC5 and the ISGylation pathway: critical modulators of the antiviral immune response[J]. Viruses, 2021, 13(6): 1102. doi: 10.3390/v13061102.

[14] Valero Y, Chaves-Pozo E, Cuesta A. Fish HERC7: phylogeny, characterization, and potential implications for antiviral immunity in European Sea bass[J]. Int J Mol Sci, 2024, 25(14): 7751. doi:10.3390/ijms25147751.

[15] Guenther M G, Frampton G M, Soldner F, et al. Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells[J]. Cell Stem Cell, 2010, 7(2): 249-57. doi:10.1016/j.stem.2010.06.015.

[16] Roth R B, Hevezi P, Lee J, et al. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS[J]. Neurogenetics, 2006, 7(2): 67-80. doi: 10.1007/s10048-006-0032-6.

[17] Lawrence B P, Denison M S, Novak H, et al. Activation of the aryl hydrocarbon receptor is essential for mediating the anti-inflammatory effects of a novel low-molecular-weight compound[J]. Blood, 2008, 112(4): 1158-65. doi:10.1182/blood-2007-08-109645.

[18] Coit P, Renauer P, Jeffries M A, et al. Renal involvement in lupus is characterized by unique DNA methylation changes in naive $CD4^+$ T cells[J]. J Autoimmun, 2015, 61: 29-35. doi:10.1016/j.jaut.2015.05.003.

[19] Gao J, Li C, Li W, et al. Increased UBE2L6 regulated by type 1 interferon as potential marker in TB[J]. J Cell Mol Med, 2021, 25(24): 11232-43. doi:10.1111/jcmm.17046.

[20] Orfali N, Shan-Krauer D, O'Donovan T R, et al. Inhibition of UBE2L6 attenuates ISGylation and impedes ATRA-induced differentiation of leukemic cells[J]. Mol Oncol, 2020, 14(6): 1297

–309. doi:10.1002/1878–0261.12614.

[21] Zhu H, Hou P, Chu F, et al. PBLD promotes IRF3 mediated the type I interferon (IFN-I) response and apoptosis to inhibit viral replication[J]. Cell Death Dis, 2024, 15(10): 727. doi:10.1038/s41419–024–07083–w.

[22] Wang J, Zheng H, Dong C, et al. Human OTUD6B positively regulates type I IFN antiviral innate immune responses by deubiquitinating and stabilizing IRF3[J]. mBio, 2023, 14(5) doi:10.1128/mbio.00332–23.

[23] AL Hamrashdi M, Brady G. Regulation of IRF3 activation in human antiviral signaling pathways[J]. Biochem Pharmacol, 2022, 200: 115026. doi:10.1016/j.bcp.2022.115026.

[24] Fan S, Popli S, Chakravarty S, et al. Non-transcriptional IRF7 interacts with NF-κB to inhibit viral inflammation[J]. J Biol Chem, 2024, 300(4): 107200. doi:10.1016/j.jbc.2024.107200.

[25] Shi H X, Yang K, Liu X, et al. Positive regulation of interferon regulatory factor 3 activation by Herc5 *via* ISG15 modification[J]. Mol Cell Biol, 2010, 30(10): 2424–36. doi:10.1128/MCB.01466–09.

Machine learning combined with bioinformatics to explore biomarkers associated with systemic lupus erythematosus diagnosis

Tang Ran[1,2], Jiang Gege[1,2], Meng Xiangwen[1,2], Cai Zheng[1,2], Jin Li[3], Xiang Nan[3], Zhang Min[3], Jia Xiaoyi[1,2]

[[1]*School of Pharmacy, Anhui University of Chinese Medicine, Hefei* 230012;
[2]*Anhui Province Key Laboratory of Bioactive Natural Products, Hefei* 230012; [3]*Dept of Rheumatology and Immunology, The First Affiliated Hospital of USTC(Anhui Provincial Hospital), Hefei* 230001]

Abstract ***Objective*** To predict and screen potential biomarkers of systemic lupus eythematosus (SLE) based on machine learning algorithms and structural biology, and to reveal their mechanisms of action and to provide new targets for disease diagnosis and treatment. ***Methods*** Four machine learning algorithms, random forest (RF), eXtreme gradient boosting (XGBoost), support vector machine (SVM), least absolute shrinkage and selection operator (LASSO), were used to analyze the gene expression data of SLE patients in GEO (datasets: GSE121239 and GSE11907) to analyze the gene expression data of SLE patients and screen key markers. Peripheral blood single nucleated cells (PBMCs) from SLE patients were collected and RT-qPCR was used to detect differential gene expression levels. Subsequently, GSEA enrichment analysis was used to identify biomarker-related pathways. CIBERSORT immune infiltration analysis and protein interactions network were applied to calculate the sample immune cell infiltration abundance. Single-cell data were analyzed for gene expression specificity in immune cells. Interaction relationships in combination with AlphaFold3 (AF3) were predicted. ***Results*** Multiple algorithms were screened together to identify the unique marker gene HERC5, and expression analysis of multiple datasets showed that HERC5 was highly expressed in SLE compared to the normal group ($P<0.05$), and RT-qPCR verified the same trend ($P=0.006\ 2$). Functional enrichment analysis identified the major pathway promoted by HERC5 in SLE as the interferon receptor signalling pathway ($P<0.05$). Immune infiltration analysis showed that HERC5 was closely associated with immune cells (Neutrophils: $r=0.39$, $P<0.05$; Memory B cells: $r=0.33$, $P<0.05$; Activated dendritic cell: $r=0.52$, $P<0.05$). Most HERC5-related interacting proteins were associated with SLE, and potential transcription factors of HERC5 and its related genes were also significantly associated with immune responses. ***Conclusion*** The HERC5 gene is an important biomarker for SLE, which upregulates the interferon pathway to promote SLE progression and provides a new target for SLE diagnosis and treatment.

Key words systemic lupus erythematosus; machine learning; bioinformatics; HERC5; interferon pathway; biomarker

Fund programs National Natural Science Foundation of China (No. 82074090); Natural Science Research Project of Anhui Educational Committee (Nos. 2024AH052061, 2024AH040154)

Corresponding authors Zhang Min, E-mail: doczhangmin@ustc.edu.cn; Jia Xiaoyi, E-mail: jiaxy@ahtcm.edu.cn