DOI: 10.19338/j.issn.1672-2019.2024.07.002

·论　著·

生物信息学联合机器学习筛选幽门螺杆菌相关萎缩性胃炎的生物标志物

卜凡靖

（滨州市第二人民医院 消化内科，山东 滨州 256800）

摘要：目的 利用加权基因共表达网络分析（WGCNA）、机器学习算法筛选幽门螺杆菌相关萎缩性胃炎（HPAG）潜在的生物标志物。**方法** 下载基因表达数据库中包含HPAG和无幽门螺杆菌感染（nonHP）的胃组织转录组数据进行差异分析，对差异表达基因（DEGs）进行基因集富集分析（GSEA）。整合WGCNA结果和DEGs，筛选HPAG相关基因。利用最小绝对收缩和选择算子（LASSO）、支持向量机递归特征消除（SVM-RFE）和随机森林（RF）等机器学习方法筛选HPAG的潜在生物标志物，提取生物标志物的表达量进行组间比较。**结果** 共获得213个DEGs，主要富集在胆固醇代谢、脂肪的消化吸收等信号通路。机器学习算法筛选出AF的潜在生物标志物S100钙结合蛋白G（S100G）。HPAG样本中S100G表达水平高于nonHP样本。**结论** HPAG发病涉及胆固醇代谢、脂肪的消化吸收等信号通路，S100G在HPAG胃组织中表达显著增高，可能成为HPAG治疗的新靶点。

关键词：萎缩性胃炎；幽门螺杆菌；加权基因共表达网络分析；机器学习；生物标志物

中图分类号：R735.2；R975

Bioinformatics combined with machine learning for screening biomarkers in *Helicobacter pylori*-associated atrophic gastritis

BU Fanjing

(Department of Gastroenterology, Binzhou Second People's Hospital, Binzhou, Shandong 256800, China)

Abstract:【Objective】 To screen potential biomarkers of *Helicobacter pylori*-associated atrophic gastritis (HPAG) using weighted gene co-expression network analysis (WGCNA), and machine learning algorithms. **【Methods】** To download the transcriptomic data of gastric tissues containing HPAG and non-*Helicobacter pylori* (nonHP) infection was from gene expression databases for differential analysis, and perform gene set enrichment analysis (GSEA) on differentially expressed genes (DEGs). WGCNA results and DEGs were integrated to screen HPAG-related genes. Machine learning methods such as least absolute shrinkage and selection operator (LASSO), support vector machine recursive feature elimination (SVM-RFE) and random forest (RF) were utilized to screen potential biomarkers for HPAG, and biomarker expressions were extracted for intergroup comparison. **【Results】** A total of 213 DEGs were obtained, which were mainly enriched in signaling pathways such as cholesterol metabolism, digestion and absorption of fat. A machine learning algorithm screened the potential biomarker of AF, S100 calcium-binding protein G (S100G). The expression level of S100G was higher in HPAG samples than in nonHP samples. **【Conclusion】** HPAG pathogenesis involves cholesterol metabolism, digestion and absorption of fat, and other signaling pathways. S100G expression was significantly increased in HPAG gastric tissues, which may become a new target for HPAG treatment.

Keywords: atrophic gastritis; *Helicobacter pylori*; weighted gene co-expression network analysis; machine learning; biomarkers

萎缩性胃炎（atrophic gastritis, AG）是临床常见的消化系统疾病，其病例特点包括胃黏膜上皮损害、固有腺体减少、伴或不伴肠腺化生和（或）假幽门腺化生等[1]。AG患者一般无特异性临床表现，主要临床症状包括腹胀、疼痛、食欲不振、反酸等[2]。幽门螺杆菌（*Helicobacter pylori*, HP）

收稿日期：2024–03–01

感染是 AG 最常见的病因[3]。据统计，AG 年发病率约为 0%~10.9%，由于检测方法存在差异，AG 发生率的报道存在较大差异[4]。肠型胃癌是最常见的胃癌类型，其发生规律为正常胃黏膜→慢性非萎缩性胃炎→AG→胃黏膜肠上皮化生→胃黏膜不典型增生→胃黏膜癌变，根除 HP 是治疗 AG 的首要措施，可部分逆转胃黏膜萎缩，降胃癌的发生风险[5]。HP 感染后进展为 AG 的机制目前尚未完全明确，目前也没有可根治 AG 的有效药物，明确 HP 感染后 AG 的发病机制具有重要临床意义。本研究拟通过加权基因共表达网络分析（weighted gene co-expression network analysis, WGCNA）和机器学习挖掘 HP 感染后 AG 发生的内在机制，挖掘潜在的生物标志物，为其机制研究提供参考。

1 资料与方法

1.1 数据来源

从基因表达数据库（GEO，http：//www.ncbi.nlm.nih.gov/geo）获得了 HP 感染后 CAG 的基因表达芯片数据集 GSE27411[6]。GSE27411 数据集共包含 6 例 HP 感染的 AG 胃组织样本（HPAG）和 6 例无 HP 感染（nonHP）的胃组织样本测序数据。

1.2 差异表达基因（differential expressed genes, DEGs）筛选和富集分析

用“limma”R 包对 GSE27411 数据集中 HPAG 和 nonHP 样本测序数据进行标准化处理，以|logFC|>1、矫正后 P 值<0.05 为条件，筛选得到差异表达基因（differentially expressed genes, DEGs），使用火山图和热图进行可视化。使用“clusterProfiler”R 包对相关基因进行基因集富集分析（gene set enrichment analysis, GSEA）。

1.3 加权基因共表达网络分析（weighted gene co-expression network analysis, WGCNA）和 HPAG 相关 DEGs 筛选

利用“WGCNA”R 包构建共表达网络，筛选最佳软阈值，计算拓扑重叠矩阵（topological overlap matrix, TOM）。使用“hclust”函数进行层次聚类，根据拓扑重叠异度（1-TOM）对基因进行模块化处理，利用“Dynamic Tree Cut”和“Module Membership”函数筛选与 HPAG 病变相关的基因模块。将 GSE27411 数据集样本 DEGs 和 WGCNA 筛选获得的与 HPAG 相关的基因模块取交集，获得 HPAG 相关 DEGs。

1.4 机器学习筛选 HPAG 生物标志物

使用“randomForest”R 包、“glmnet”R 包和“e1071”R 包[5-7]对 GSE27411 数据集中 HPAG 和 nonHP 样本测序数据进行最小绝对收缩和选择算子（least absolute shrinkage and selection operator, LASSO）、支持向量机递归特征消除（support vector machine recursive feature elimination, SVM-RFE）和随机森林（random forest, RF）分析，将三种模型筛选获得的基因取交集，筛选 HPAG 生物标志物。提取 HPAG 生物标志物在 GSE27411 数据集各组样本中的表达量，进行差异分析。

2 结果

2.1 GSE27411 数据集 DEGs 筛选

使用“limma”R 包对 GSE27411 数据集中 HPAG 和 nonHP 胃组织样本测序数据进行标准化处理并筛选 DEGs（图 1A），共获得 20 590 个表达量大于 0 的基因，筛选获得 213 个 DEGs，其中差异表达上调基因 201 个，差异表达下调基因 12 个。绘制 DEGs 火山图和热图（图 1B、图 1C）。

2.2 DEGs GSEA 分析

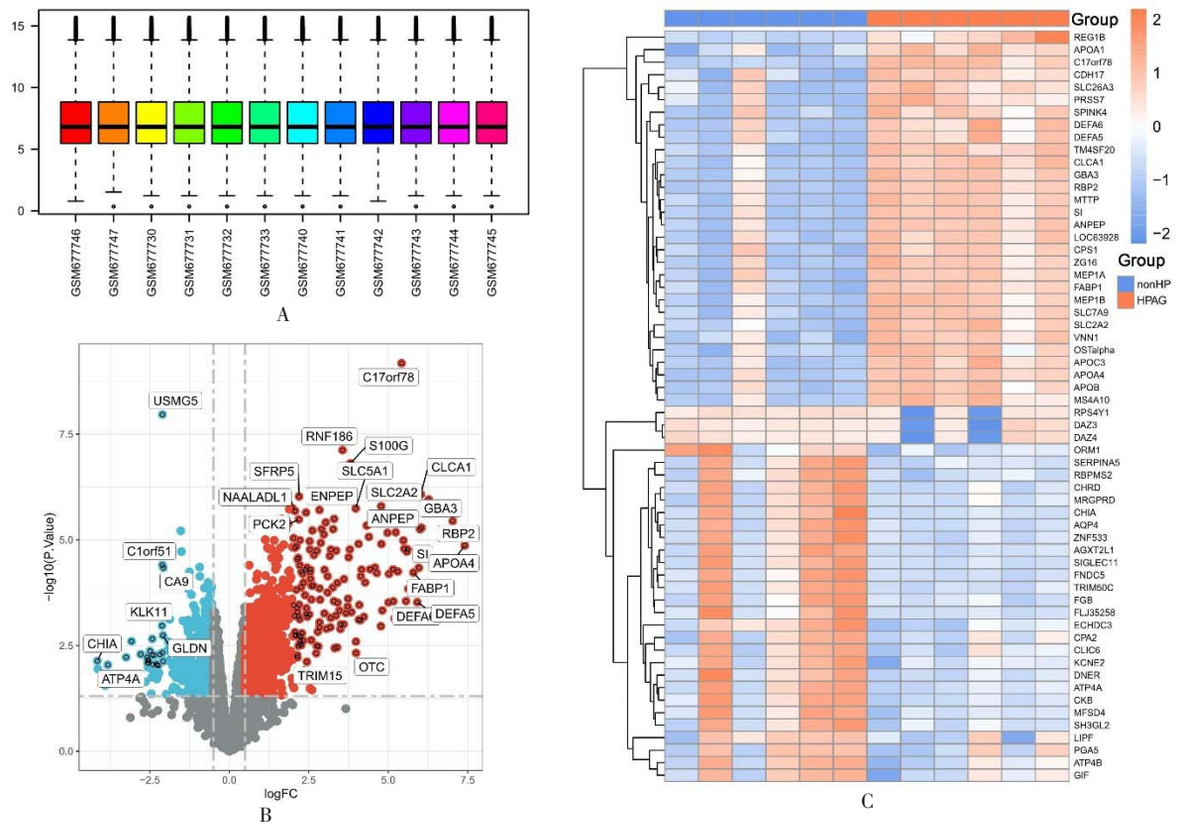
富集分析结果显示，差异表达上调 DEGs 显著富集于碳水化合物的消化和吸收、胆固醇代谢、脂肪的消化吸收、维生素的消化吸收等信号通路（图 2A），差异表达下调 DEGs 显著富集于昼夜节律、胃酸分泌、TGF-B 信号通路等通路上（图 2B）。

2.3 WGCNA 和 HPAG 相关基因筛选

使用“WGCNA”包的“pickSoftThreshold”函数对表达量大于 0 的基因进行筛选，将软阈值设为 8，建立无尺度网络（图 3A）。将阈值设为 0.4，最小模块基因数设为 100，共聚类出 12 个模块（图 3B）。模块–性状关联分析显示，蓝色和浅黄色模块与 HPAG 显著相关（图 3C）。将 DEGs 和 WGCNA 筛选获得的与 HPAG 病变相关的基因模块取交集，获得 189 个 HPAG 相关基因（图 3D）。

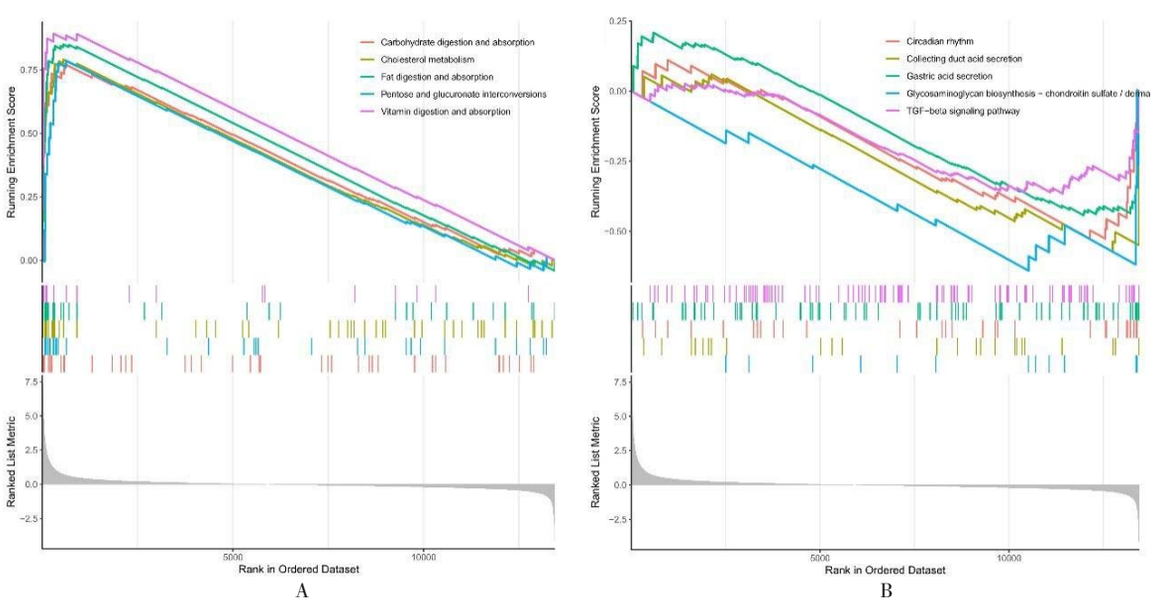
2.4 HPAG 生物标志物识别

使用 LASSO 回归识别出 8 个关键的生物标志物（图 4A），使用 SVM-RFE 法识别出 1 个关键生物标志物（图 4B），使用 RF 算法识别出 4 个关键生物标志物（图 4C），三种方法筛选出 1 个关键生物标志物：S100 钙结合蛋白 G（S100 calcium-binding protein G, S100G）（图 4D）。差异分析结果提示，HPAG 胃组织样本中 S100G 表达量高于 nonHP 胃组织样本，差异有统计学意义（P=0.0022）（图 5）。



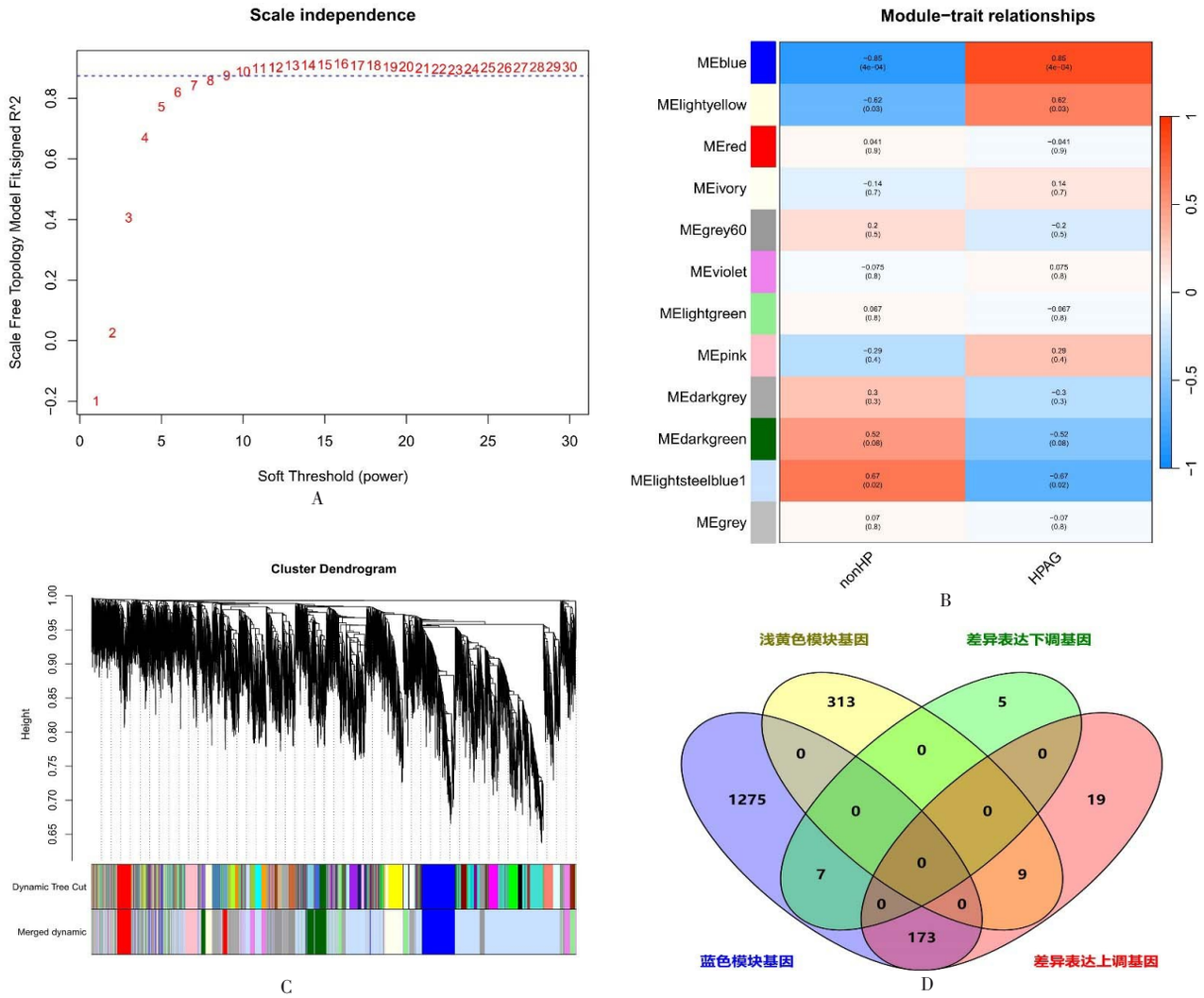
A：标准化处理后数据柱状图；B：DEGs 火山图；C：DEGs 热图。

图 1　GSE27411 数据集 DEGs 筛选



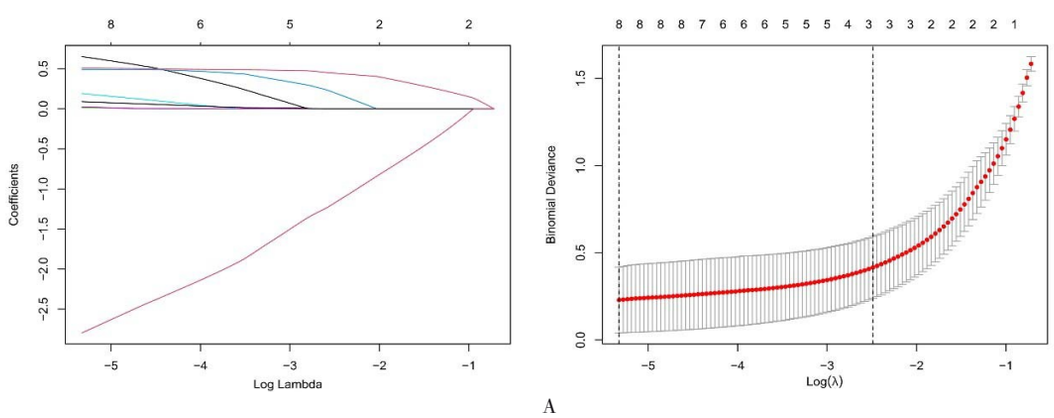
A：差异表达上调 DEGs GSEA 富集；B：差异表达下调 DEGs GSEA 富集。

图 2　DEGs GSEA 富集



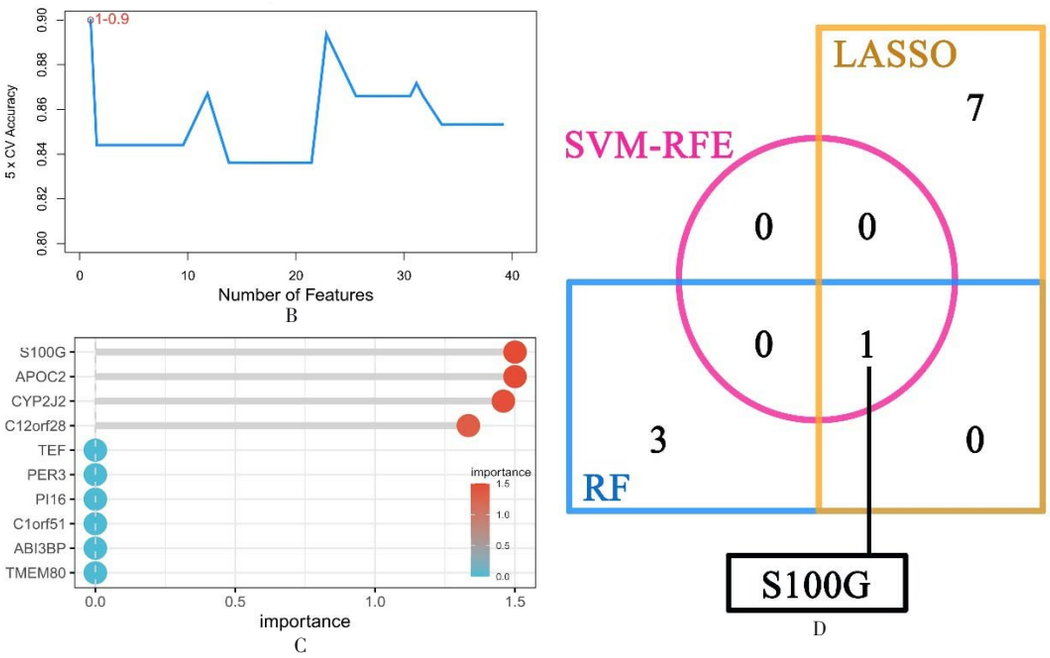
A：最佳软阈值筛选；B：基因模块聚类；C：模块–性状关联分析热图；D：WGCNA 与 DEGs 基因韦恩图。

图 3　WGCNA 和 HPAG 相关基因筛选



A：LASSO 筛选生物标志物。

图 4　HPAG 生物标志物识别



B：SVM-RFE筛选生物标志物；C：RF筛选的生物标志物重要性棒棒糖图；D：机器学习筛选生物标志物韦恩图。

续图 4　HPAG 生物标志物识别

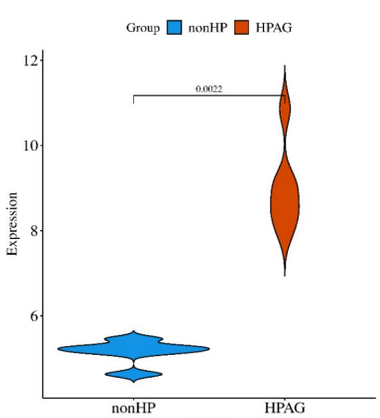


图 5　S100G 差异表达分析

3　讨论

本研究对包含 HPAG 和 nonHP 胃组织样本测序数据进行分析，并利用 WGCNA、SVM-RFE、RF、LASSO 等机器学习方法筛选与 HPAG 相关的生物标志物，结果显示，HPAG 相关的 DEGs 以差异表达上调为主，显著富集于胆固醇代谢、脂肪的消化吸收等信号通路上，S100G 是 HPAG 潜在的生物标志物。

既往研究证实，HP 感染可诱导轻度炎症，在炎症微环境下，肿瘤坏死因子、白细胞介素及多种炎症相关的黏附分子表达上调，诱导脂质过氧化以及低密度脂蛋白胆固醇氧化修饰影响脂质代谢[7]。一项横断面研究证实，约 71.8% 的 HP 感染患者存在血脂异常，与对照组患者比较，HP 感染患者低密度脂蛋白胆固醇、甘油三酯、总胆固醇水平升高，HP 感染是血脂异常的独立危险因素[8]。AG 与 HP 感染密切相关，有研究发现，随着 HP 感染时间延长，胃黏膜的炎症和肠上皮化生程度越严重[9]。既往研究发现，AG 的发生发展与脂质过氧化物作用密切相关[10]。黄健等[11]采用倾向性匹配法分析发现，AG 患者甘油三酯水平高于非萎缩性胃炎患者。日本一项单中心研究发现，与根除 HP 治疗前相比，根除 HP 治疗后患者体重、体重指数和肥胖指数显著增加，白细胞和血小板计数显著降低，高密度脂蛋白胆固醇水平显著升高，高密度脂蛋白胆固醇/低密度脂蛋白胆固醇比率显著降低[12]。HP 以及 AG 与脂肪代谢的相关性研究较少，王宏星[13]统计国人膳食习惯、营养元素摄入与 AG 的相关性发现，单纯高脂肪摄入是 AG 发病的危险因素。

S100G 是 S100 蛋白家族成员，S100 是一组具有多种生物学功能的低分子质量的钙结合蛋白，在调节钙信号转导、炎症、细胞生长与分化、细胞骨架动态以及细胞间交流等过程中发挥作用，与细胞的屏障功能和通透性紧密相关[14-15]。研究证实，S100 蛋白家族成员可影响消化道黏膜的通透性，参与溃疡性结肠炎发病[16]。另有研究发现，

HP 感染阳性的患者 S100A8、S100A9 表达阳性率明显高于 HP 感染阴性的患者[17]。但也有研究发现，S100G 可通过抑制核转录因子 Kappa B 活化来抑制单核细胞趋化蛋白-1 的产生发挥抗炎作用[18]。S100G 与 AG 和 HP 感染的相关性尚不明确，结合上述研究和本文的分析结果，笔者分析，S100G 可能通过调控胆固醇、脂肪代谢等途径，参与 HP 感染和 AG 的发生、发展。

综上所述，本研究通过生物信息学分析和机器学习算法，筛选出 HP 感染 AG 相关的生物标志物 S100G，S100G 参与 HP 感染 AG 的途径可能包括胆固醇、脂肪代谢等。

参 考 文 献

[1] 李军祥, 陈誩, 吕宾, 等. 慢性萎缩性胃炎中西医结合诊疗共识意见(2017 年)[J]. 中国中西医结合消化杂志, 2018, 26(2): 121-131.

[2] 王亚杰, 国嵩, 杨洋, 等. 慢性萎缩性胃炎的流行病学及其危险因素分析[J]. 中国中西医结合消化杂志, 2019, 27(11): 874-878.

[3] 石振旺, 方东, 鲍德明, 等. 胃癌高发地区慢性萎缩性胃炎血清幽门螺杆菌抗体分型情况以及癌变风险的差异[J]. 安徽医药, 2023, 27(2): 332-336.

[4] HOLLECZEK B, SCHÖTTKER B, BRENNER H. *Helicobacter pylori* infection, chronic atrophic gastritis and risk of stomach and esophagus cancer: results from the prospective population-based ESTHER cohort study[J]. Int J Cancer, 2020, 146(10): 2773-2783.

[5] 国家消化系疾病临床医学研究中心(上海), 国家消化道早癌防治中心联盟, 中华医学会消化病学分会幽门螺杆菌学组, 等. 中国胃黏膜癌前状态和癌前病变的处理策略专家共识(2020 年)[J]. 中华消化杂志, 2020, 40(11): 731-741.

[6] NOOKAEW I, THORELL K, WORAH K, et al. Transcriptome signatures in *Helicobacter pylori*-infected mucosa identifies acidic mammalian chitinase loss as a corpus atrophy marker[J]. BMC Med Genomics, 2013, 6: 41.

[7] CHEN Y, YOU NN, YANG CY, et al. *Helicobacter pylori* infection increases the risk of carotid plaque formation: clinical samples combined with bioinformatics analysis[J]. Heliyon, 2023, 9(9): e20037.

[8] NIGATIE M, MELAK T, ASMELASH D, et al. Dyslipidemia and its associated factors among *Helicobacter pylori*-infected patients attending at university of Gondar comprehensive specialized hospital, Gondar, north-west Ethiopia: a comparative cross-sectional study[J]. J Multidiscip Healthc, 2022, 15: 1481-1491.

[9] 李温静, 董全江, 于新娟, 等. 幽门螺杆菌感染与肠上皮化生的相关性[J]. 青岛大学学报(医学版), 2020, 56(6): 687-690.

[10] 吕涛, 刘皓, 魏睦新. 益生菌联合化痰消瘀汤治疗慢性萎缩性胃炎疗效及对脂质过氧化损伤指标的影响[J]. 现代中西医结合杂志, 2017, 26(21): 2281-2283, 2297.

[11] 黄健, 李定富, 李玉萍. 基于倾向性评分匹配探索慢性萎缩性胃炎与血脂的关系[J]. 国际检验医学杂志, 2023, 44(12): 1464-1467.

[12] IWAI N, OKUDA T, OKA K, et al. *Helicobacter pylori* eradication increases the serum high density lipoprotein cholesterol level in the infected patients with chronic gastritis: a single-center observational study[J]. PLoS One, 2019, 14(8): e0221349.

[13] 王宏星. 萎缩性胃炎患者膳食与营养情况调查分析[J]. 公共卫生与预防医学, 2019, 30(1): 126-129.

[14] ZHOU Y, ZHA YW, YANG YQ, et al. S100 proteins in cardiovascular diseases[J]. Mol Med, 2023, 29(1): 68.

[15] NOACK M, MIOSSEC P. Heterogeneous effects of S100 proteins during cell interactions between immune cells and stromal cells from synovium or skin[J]. Clin Exp Immunol, 2023, 212(3): 276-284.

[16] 檀飞飞, 周中银. 黏膜通透性相关因子 CK8、S100、E-cadherin 在溃疡性结肠炎中的表达及作用[J]. 临床消化病杂志, 2023, 35(6): 443-446.

[17] 卜佳, 孙曼銮, 吕科, 等. 不同 Hp 感染的老年慢性胃炎患者 S100 蛋白表达及意义[J]. 中国老年学杂志, 2023, 43(22): 5446-5449.

[18] ISHIGURO K, WATANABE O, NAKAMURA M, et al. S100G expression and function in fibroblasts on colitis induction[J]. Int Immunopharmacol, 2016, 39: 92-96.

（龚仪 编辑）