克罗恩病和胰腺炎中的共同基因特征及生物学机制鉴定

胡天媛[1] 姜政[2*]
1. 重庆医科大学，中国·重庆 400015
2. 重庆医科大学第一附属医院，中国·重庆 400016

摘 要：目的：从基因综合表达数据库（GEO）中获取克罗恩病（CD）和胰腺炎的基因表达数据，通过筛选差异基因和核心表型基因并进行功能分析，探索克罗恩病和胰腺炎的潜在发病机制。**方法：**在 GEO 数据库中检索 CD 和胰腺炎的全血样本基因表达数据集，下载合适的数据集 GSE3365、GSE194331 的原始数据，筛选 CD 和胰腺炎的共享差异表达基因（DEGs），并通过使加权基因共表达网络（WGCNA）筛选共同核心表型基因，在此基础上对共享基因进行功能富集分析、蛋白质互作网络（PPI）和三种机器学习算法分析。**结果：**筛选出 32 个 DEGs 及 44 个共享模块基因，功能富集分析强调了炎症趋化因子和细胞因子在这两种疾病机制中的重要性。PPI 分析识别出 9 个关键枢纽基因。最终，通过三种机器学习算法确定了 4 具有潜在诊断价值的共享基因 PDGFB、IL10、IL6、FGF13。**结论：**炎症反应的激活和调节、炎症后纤维化、脂质代谢的生物过程改变与 CD 并发胰腺炎的病程有关键作用，IL6、PDGFB、IL10、FGF13 可认为是其预测因子，JAK-STAT 信号通路可能为其关键通路。
关键词：克罗恩病；胰腺炎；生物标志物；生物信息学

Identification of Common Gene Characteristics and Biological Mechanisms in Crohn's Disease and Pancreatitis

Tianyuan Hu[1] Zheng Jiang[2*]
1. Chongqing Medical University, Chongqing, 400015, China
2. The First Affiliated Hospital of Chongqing Medical University, Chongqing, 400016, China

Abstract: Objective: To obtain gene expression data of Crohn's disease (CD) and pancreatitis from the Gene Omnibus Expression Database (GEO), and explore the potential pathogenesis of CD and pancreatitis by screening differential genes and core phenotype genes and conducting functional analysis. **Method:** Retrieve the gene expression datasets of whole blood samples of CD and pancreatitis from GEO database, download the appropriate raw data of GSE3365 and GSE194331, screen for shared differentially expressed genes (DEGs) of CD and pancreatitis, and use weighted gene co expression network (WGCNA) to screen for common core phenotype genes. Based on this, perform functional enrichment analysis, protein-protein interaction network (PPI), and three machine learning algorithms analysis on the shared genes. **Result:** 32 DEGs and 44 shared module genes were screened, and functional enrichment analysis emphasized the importance of inflammatory chemokines and cytokines in these two disease mechanisms. PPI analysis identified 9 key hub genes. Finally, four shared genes with potential diagnostic value, PDGFB, were identified through three machine learning algorithms IL10, IL6, FGF13. **Conclusion:** The activation and regulation of inflammatory response, post inflammatory fibrosis, and changes in biological processes of lipid metabolism play a key role in the course of CD complicated pancreatitis. IL6, PDGFB, IL10, and FGF13 can be considered as predictive factors, and the JAK-STAT signaling pathway may be a key pathway.
Keywords: Crohn's disease; pancreatitis; biomarkers; bioinformatics

1 绪论

炎症性肠病（IBD）是一种免疫介导的慢性非特异性肠道炎症，临床上包括溃疡性结肠炎及克罗恩病。近年来，炎症性肠病的病例数愈发增多，中国的 IBD 患病率也在逐渐增高。一项基于 GBD 数据库的流行病学研究显示：1990—2019 年中国 IBD 的标化发病率呈逐年上升趋势，疾病负担重，对中国来说是一个重大的公共卫生挑战[1]。且长期以来，IBD 的具体发病机制未能得到明确的解答，使得许多 IBD 患者无法得到及时的诊断与治疗。

现有的多项研究已表明，IBD 是一种系统性疾病，其临床表现除了病变引起的消化道症状外，常可伴有多系统受累的肠外表现（EIM），包括关节、皮肤、血管、心肺及肝胆胰系统等[2]。其中，胰腺损伤为 IBD 的一种常见肠外表现。肠外表现与 IBD 患者的临床治疗及其预后有着重要的相关性，对 IBD 相关胰腺损伤发病机制的研究有助于为 IBD 相

关病变寻找预测因子，并分析可能的治疗靶点，为患者提供更加个性化的治疗。本研究旨在通过生物信息学分析方法及相关分析工具，进一步研究 CD 与胰腺炎之间的共同生物学机制。

2 克罗恩病与急性胰腺炎共病机制的生物信息学分析

2.1 数据来源

通过 GEO 数据库搜索并下载两个 mRNA 数据集，“GSE3365”和“GSE194331”。“GSE3365”包括 59 个克罗恩病患者及 42 个正常对照者共 101 人血液样本基因表达数据，“GSE194331”包含 87 个胰腺炎患者及 32 个正常对照者共 119 人的血液样本基因表达数据。

2.2 差异基因的筛选

在 CD 数据集中，与正常人对照组相比，从 GSE3365 和 GSE194331 两个数据集中分别筛选出差异基因 1322 个和 4180 个差异基因，并分别绘制火山图，|log2FC(foldchange)| ＞ 0.5，P.Value ＜ 0.05，如图 1A、1B 所示。其中，包括 10 个共有上调差异基因和 22 个共有下调差异基因，如图 1C、1D 所示。共有上调差异基因为：ADAMTS6、MAP2、PDGFB、ASIC1、POF1B、FOSB、OLFM1、IL6、HSPG2、GPR183，共有下调差异基因为：SEC62、PPP1R12A、KLF6、APBB1IP、NME8、CDADC1、USP10、ELF2、WBP4、RAI2、SORL1、PPIG、PGGHG、UGCG、ADAM8、MPPE1、TSC22D3、ABHD3、IGFBP7、H1-4、EHBP1L1、EBLN2。

2.3 差异基因的 GO 富集分析及 KEGG 分析

对共同差异基因进行基因本体论（GO）分析。对共有上调基因分析得到 GO 分类列表如图 2A 所示，其生物学过程（BP）分析表明，共有上调差异基因在免疫反应、炎症细胞趋化、血小板活化调节、上皮间质转化等过程有关；细胞组成（CC）分析表明，共有上调基因主要与树突、血小板衍生生长因子复合体、白介素 -6 受体复合物等细胞组分有关；分子功能（MF）分析表明，共有上调差异基因与生长因子受体结合、生长因子活性、白介素 -6 受体结合、酸敏感离子通道活性、NAPDH 氧化酶激活剂活性等过程有关。对共有下调基因分析得到 GO 分类列表如图 2B 所示，其分子功能分析表明共同下调基因与神经酰胺葡萄糖基转移酶活性、蛋白质 - 葡糖基半乳糖基羟赖氨酸葡萄糖苷酶活性、GPI- 甘露糖乙醇胺磷酸二酯酶活性、二氢神经酰胺葡萄糖基转移酶活性等功能有关。

对共同差异基因进行京都基因与基因组百科全书（KEGG）分析。对共有上调基因分析得到 KEGG 通路列表如图 2C 所示，可看到共有上调基因富集显著的通路有 EGFR 酪氨酸激酶抑制剂耐药性、IL-17 信号通路、JAK-STAT 信号通路、PI3K-Akt 信号通路等。同时，对共有下调基因行 KEGG 分析，通路列表如图 2D 所示，可看到共有下调基因富集显著的通路有血小板激活、蛋白质出口、鞘脂类代谢、血管平滑肌收缩、cGMP-PKG 信号通路、Rap1 信号通路、CAMP 信号通路等。

2.4 WGCNA 筛选核心表型基因

对两组数据进行加权基因共表达网络（WGCNA）分析。对于 CD 组，选择最佳软阈值为 5，如图 3A 所示。后使用邻接函数生成邻接矩阵、使用 TOM 相异度度量构建了基因的层次聚类，如图 3B 所示。分析共识别出 19 个共表达模块，P 值＜ 0.05 的模块被认为是关键模块，如图 3C 所示，black 和 brown 模块与疾病表型有着强正相关，一共包括 348 个基因，而 greenyellow、cyan、red 和 green 模块与疾病表型有着强负相关，一共包括 338 个基因。用同样的方法分析胰腺炎组，选择最佳软阈值为 5，如图 3D 所示。共识别出 13 个共表达模块，blue 模块与疾病表型有着强正相关，包括 1238 个基因，black 模块与疾病表型有着强负相关，包括 40 个基因，如图 3E、3F 所示。

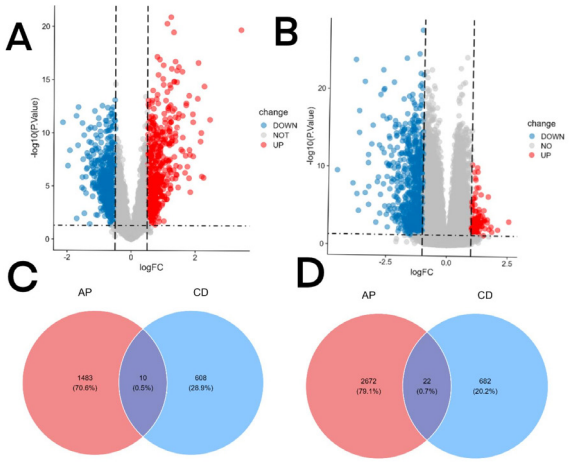


图 1（A）GSE3365 的火山图；（B）GSE194331 的火山图；上调基因为红色，下调基因为蓝色；（C）CD 和 AP 之间共享的上调差异基因；（D）CD 和 AP 之间共享的下调差异基因

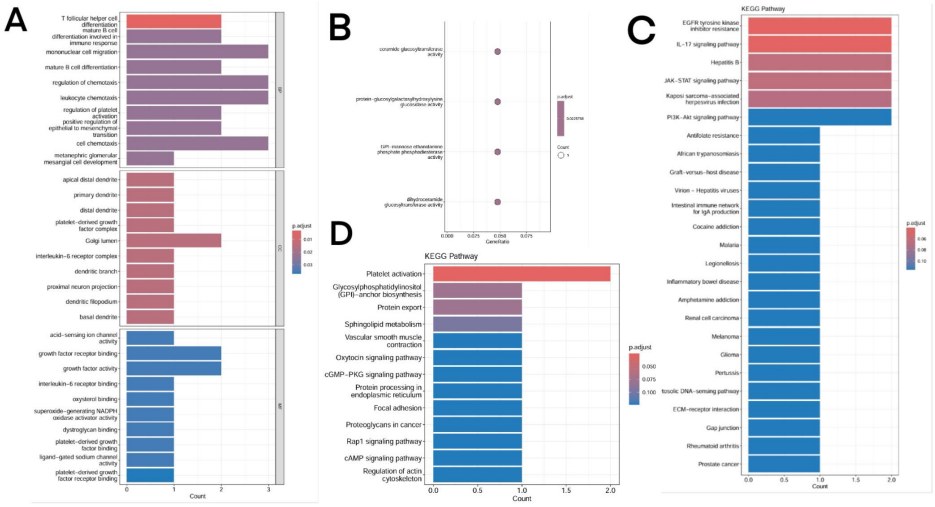
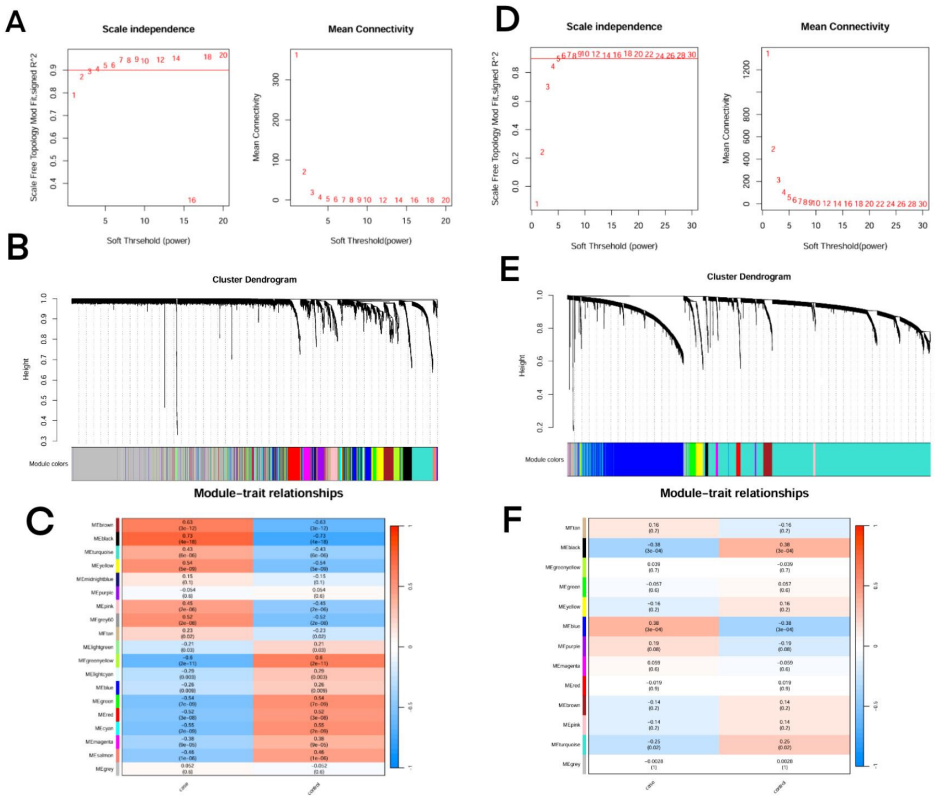


图 2（A）上调基因 GO 分析；（B）下调基因 GO 分析；（C）上调基因 KEGG 分析；（D）下调基因 KEGG 分析



（A）CD 的软阈值确定；（B）CD 模块中高度连接基因的聚类树状图；（C）CD 中模块与特征之间的关系，每个单元格中包含相关性和 P 值；（D）AP 的软阈值确定；（E）AP 模块中高度连接基因的聚类树状图；（F）AP 中模块与特征之间的关系，每个单元格中包含相关性和 P 值。

图 3 CD 和 AP 的加权基因共表达网络分析（WGCNA）

2.5 核心表型基因的 GO 富集分析及 KEGG 分析

将上述分析得出的与两组疾病发生有关的核心表型基因取交集，共得到 44 个共有核心表型基因，如图 4A 所示。这些共有核心表型基因可能与两种疾病的共同发生机制有较强的关联性。为了进一步探究两种疾病共同发生的机制，对共有基因也进行 GO 富集分析及 KEGG 分析，分析得到 GO 分类列表如图 4B 所示。BP 分析表明，共有核心表型基因与排卵、白细胞趋化性、miRNA 转录的正调控等生物过程有关。MF 分析显示，共有核心表型基因在细胞因子活性、生长因子活性、环化酶调节剂活性、补体组分 C3a 受体活性、白介素 10 受体结合等分子功能有关。对共有枢纽基因分析得到 KEGG 通路列表如图 4C 所示，共有核心表型基因富集显著的通路有细胞因子 - 细胞因子受体相互作用、补体和凝血级联、癌症中的转录失调、NF-kappa B 信号通路等。

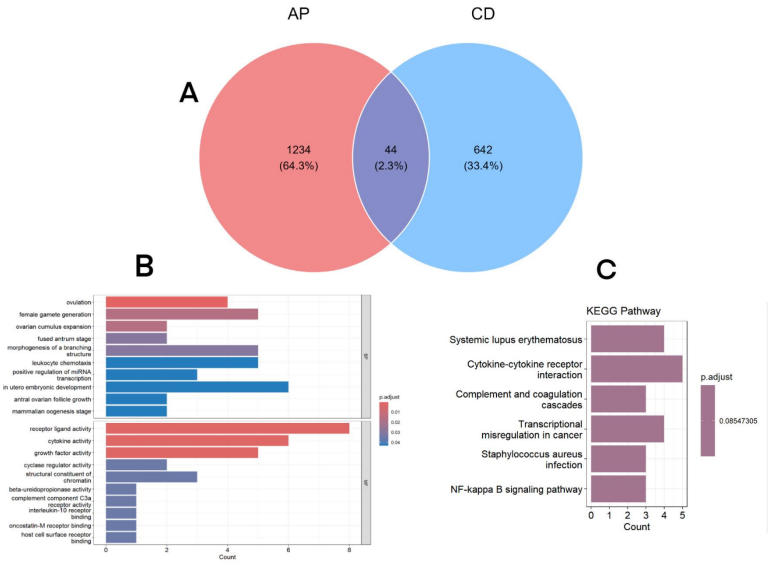


图 4（A）CD 和 AP 之间共享的模块基因；（B）共享基因的 GO 分析；（C）共享基因的 KEGG 分析

2.6 PPI（蛋白质互作网络）分析

筛选出的差异基因与核心表型基因，最终会通过蛋白质的形式表现出来。通过检索编码蛋白之间可能的潜在相互作用并构建蛋白质互作网络，可以体现出这些基因或蛋白质之间存在着怎样的相互关系，最终找到有意义的分子调节网络。首先，将共有差异基因列表上传至 STRING 数据库，选择 multiple proteins 模块，设置可信度为 0.7，隐藏未有连接的点，得到 PPI 网络图（见图 5A）。将互联网络文件导入 Cytoscape，使用软件中的 MCODE 插件寻找最密切相关的基因模块，得到一个关键子网络（见图 5B），得分为 3 分，含有 3 个节点，3 种互作关系，节点为 FOSB、KLF3、IL6。再使用 Cytoscape 自带的 7 种算法筛选互作网络中的枢纽基因。使用的算法有 MCC（最大集团中心性）、Degree（节点度数值）、Closeness（紧密度）、Radiality（径向度）、Stress（应力）、EPC（边缘渗透组件）和 MNC（邻域组件中心性）。选取算法得出的得分 TOP6 基因作为枢纽基因，通过取交集得到了 4 个关键枢纽基因：IL6、KLF6、SORL1、PDGFB。

使用同样的方法分析核心表型基因。基因列表上传至 STRING 数据库，基本设置同前，得到 PPI 网络图（见图 6A）。导入 Cytoscape，使用 MCODE 插件，得到两个关键子网络（见图 6B、图 6C）。子网络 1 得分为 3.778 分，含有 10 个节点，17 种互作关系，节点为 NAMPT、TNFAIP6、FGF13、MSR1、OSM、IL10、SOCS3、CXCL3、C3AR1、EREG。子网络 2 得分为 3 分，含有 3 个节点，3 种互作关系，节点为 H3C8、H2AC8、H2AC13。使用同样的 7 种算法寻找枢纽基因，得到的 TOP10 基因，通过取交集得到了 5 个关键枢纽基因：IL10、SOCS3、FGF13、BMP2、CXCL3。

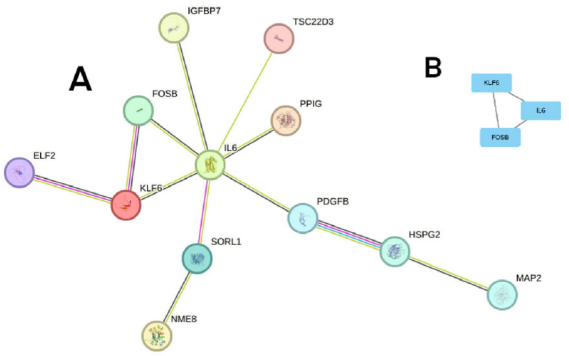


图 5（A）共享差异基因的 PPI 网络；（B）重要子网络

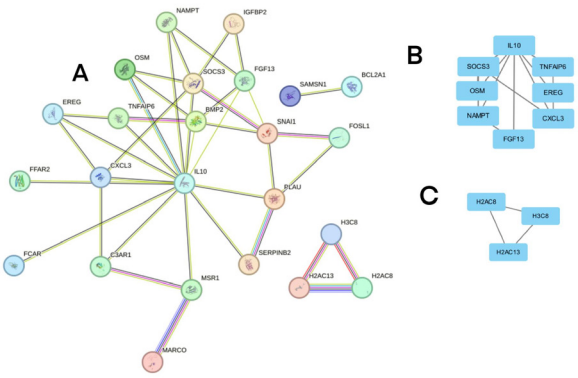


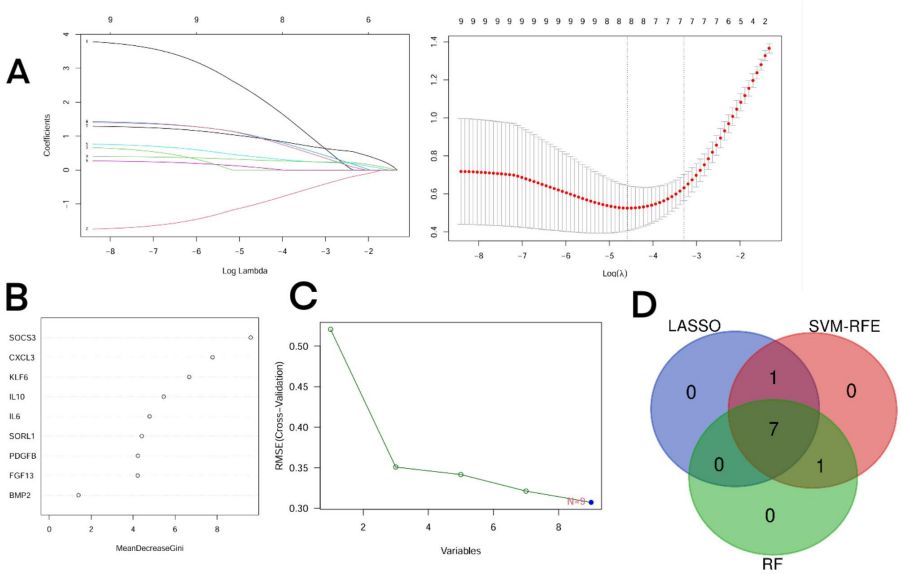
图 6（A）共享模块基因的 PPI 网络；（B）重要子网络 1；（C）重要子网络 2

2.7 机器学习筛选潜在诊断基因

通过上述步骤，已筛选出 9 个与两种疾病密切相关的关键枢纽基因。为了进一步选择具有显著疾病组和对照组分类特征值的潜在诊断基因，对上述 9 个关键枢纽基因，使用 3 种机器学习方法筛选，分别为 LASSO（Least absolute

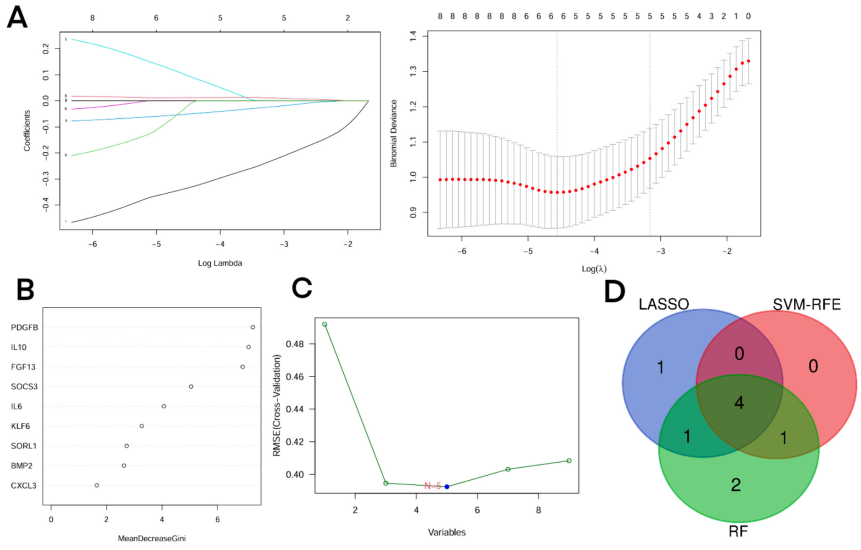
shrinkage and selection operator）、RF（Random Forest）和 SVM（Support Vector Machine）。首先对 CD 基因表达数据进行 LASSO 分析。根据 LASSO 系数分布和最佳协调参数选择图，lambda 设置为 8（见图 7A）。之后，得到 8 个非零系数的基因。然后进行随机森林分析，将上述 9 个基因输入 RF 分类器中，并在重要性量表上显示评分大于 2 的基因，得到 8 个基因（见图 7B）。最后进行 SVM 法分析，N=9 时错误率最低，故筛选出 9 个关键基因（见图 7C）。CD 组三种机器学习分析出的关键诊断基因取交集，得到 7 个共有关键基因（见图 7D），分别为 PDGFB、KLF6、IL10、SOCS3、CXCL3、IL6、FGF13。

以同样的方法将 9 个关键基因放入 AP 基因表达数据集中进行三种机器学习筛选。根据 LASSO 系数分布和最佳协调参数选择图，lambda 设置为 6（见图 8A），得到 6 个非零系数的基因。放入 RF 分类器中，在重要性量表上显示评分大于 2 的基因，得到 8 个基因（见图 8B）。在 SVM 分析方法中，N=5 时错误率最低，故筛选出 5 个关键基因（见图 8C）。将 AP 组三种机器学习分析出的关键诊断基因取交集，得到 4 个共有关键基因（见图 8D），分别为 PDGFB、IL10、IL6、FGF13。



（A）LASSO 模型系数分布图显示 CD 的 lambda 选择；（B）RF 算法判别的 CD 中评分大于 2 的基因；（C）SVM-RFE 算法选择的 9 个 CD 串扰基因；（D）韦恩图展示三种算法相交得到的 7 个 CD 诊断候选基因。

图 7 使用三种机器学习算法筛选 CD 的诊断基因



（A）LASSO 模型系数分布图显示 AP 的 lambda 选择；（B）RF 算法判别的 AP 中评分大于 2 的基因；（C）SVM-RFE 算法选择的 5 个 AP 串扰基因；（D）韦恩图展示三种算法相交得到的 4 个 CD 诊断候选基因。

图 8 使用三种机器学习算法筛选 AP 的诊断基因

最终将 CD 组与 AP 组筛选出的关键潜在诊断基因取交集，得到了 4 个共有关键基因：PDGFB、IL10、IL6、FGF13，即为本研究重点关注的生物标志物。

3 讨论与结论

IL-6（白细胞介素 -6，白介 6）是趋化因子家族的一种细胞因子，在炎症反应、免疫反应及造血作用等多个方面具有广泛的功能。IL-6 的失调与多种慢性炎症和自身免疫紊乱性疾病有关，通过由 IL-6 本身、其受体 IL-6R 和糖蛋白 130（IL-6/IL-6R/gp130）组成的六聚体复合物介导其生物学作用 [3]。IL-6 对于肠道的影响已有多方面的研究。IL-6 可参与肠道菌群功能的调节，IL-6 的表达与肠道粘膜的完整性有关。而在胰腺炎的炎症级联反应中，IL-6 也起到了关键的作用。IL-6–STAT3 通路的激活在胰腺炎发病期间进一步释放促炎因子，引发全身性的炎症反应，而 IL-6–STAT3 通路也是 IBD 治疗的一个关键靶点。综上所述，IL-6 介导的 JAK-STAT 信号通路可能为 CD 并发胰腺炎的有效治疗靶点。

PDGFB 是血小板源性生长因子（PDGF）家族的成员，与血管内皮生长因子（VEGF）家族关系相近。现有研究显示，PDGF 过表达通常与动脉粥样硬化、纤维化与恶性肿瘤有关，尤其是与恶性肿瘤中血管的形成有关。在 CD 发病过程中，巨噬细胞会通过产生 PDGF，驱动肌成纤维细胞激活和细胞外基质沉积，从而促进 CD 的纤维化进程。既往的研究发现，IBD 合并胰腺损伤的胰腺病理显示伴有间质纤维化的表现。两种疾病对 PDGFB 的共同差异表达显示出两者共有的炎症后靶器官纤维化风险，可能为 CD 合并 AP 预后不良的一种原因。

IL-10 与 IL-6 相似，是一种具有重要免疫调节功能的细胞因子，相较于 IL-6，它是一种具有强大抗炎特性的多效性细胞因子，主要由抗原呈递细胞分泌，通过激活巨噬细胞抑制炎症细胞因子如 TNF-α、IL-6 和 IL-1 的表达。IL-10 通过调节肠上皮细胞维持肠黏膜屏障稳态。但在胰腺炎病程中，早期患者血液中 IL-10 的增加被认为与胰腺炎的严重程度密切相关。在 CD 病程中，IL-10 因炎症反应表达上调，可能为并发胰腺炎的原因。

FGF13 是成纤维细胞生长因子（FGF）家族的一员，拥有调节电压门控钠通道的功能，在大脑皮层和海马体的神经元极化和迁移中起着至关重要的作用。目前关于 FGF13 和 CD、胰腺炎等炎症反应类疾病有关联的研究还鲜有提及，但有研究认为 FGF13 是造成肥胖的新候选基因，它在脂肪细胞功能中起到了重要的作用。高脂血症为 CD 与胰腺炎共同的易感因素，可能与 FGF13 的表达有所关联。

综上所述，克罗恩病和胰腺炎的共病机制受多个因子的共同作用，主要发病机制与炎症反应的激活和调节、炎症后器官的纤维化有关，IL6、PDGFB、IL10、FGF13 可认为是其预测因子，JAK-STAT 信号通路可能为其关键通路。

参考文献：

[1] 包云丽,汪哲,唐海茹,等.1990—2019年中国炎症性肠病疾病负担及变化趋势分析[J].中国全科医学,2023,26(36):4581-4586.

[2] 李秋瑾,唐佳惠,周建明,等.炎症性肠病的胰腺表现[J].胃肠病学,2020,25(1):7-12.

[3] 刘莉,赵春华,闵寒.白细胞介素-6与肠易激综合征:机制与研究进展[J/OL].协和医学杂志,1-14[2025-03-10].