

基于唾液16S rRNA基因测序和机器学习算法确定血瘀体质肺结节良恶性诊断的微生物标志物

黄春霞^{1†}, 马琼^{1†}, 钟杨¹, 曾潇¹, 何佳玮¹, 李雪珂^{1,2}, 王倩¹, 付西^{1,3}, 任益锋^{1,3}, 由凤鸣^{1,3*}

1. 成都中医药大学附属医院, 代谢性疾病中医药调控四川省重点实验室, 成都 610075

2. 成都中医药大学肿瘤学教研室, 成都 610075

3. 成都中医药大学肿瘤研究所, 成都 610075

† 同等贡献

* 联系人, E-mail: yfmdoc@163.com

2024-12-13 收稿, 2025-02-20 修回, 2025-03-11 接受, 2025-03-19 网络版发表

国家自然科学基金(82405359)、四川省中医药管理局项目(2023ZD06)和四川省科技学术厅项目(2024NSFSC1868)资助

摘要 血瘀质是肺癌的高风险体质类型, 在肺癌发生的早期阶段——肺结节, 特别关注血瘀质人群具有重要意义. 本研究旨在基于口腔微生物特征构建血瘀质肺结节良恶性诊断模型, 识别血瘀质肺结节早期恶性诊断的新型非侵入性生物标志物, 并了解口腔菌群的变化在肺结节恶性演进中的潜在作用. 研究收集了肺结节患者的唾液样本及中医体质信息, 并将血瘀质良性($n=136$)和恶性肺结节($n=73$)患者作为研究对象. 运用16S rRNA基因测序技术和7种机器学习算法(logistic regression、support vector machine、multi-layer perceptron、naïve bayes、random forest、gradient boosting decision trees、light gradient boosting machine (lightGBM)), 揭示血瘀质良恶性肺结节患者的口腔微生物组特征, 并确定区分血瘀质良恶性肺结节的最佳预测模型和关键微生物标志物. 最后, 基于PICRUST2菌群功能预测分析, 探索口腔微生物群在肺结节恶性演进中的潜在作用机制. 研究发现, 与血瘀质良性肺结节相比, 恶性肺结节患者的唾液微生物 α 和 β 多样性显著降低($P<0.05$), 且物种组成更少. LightGBM是预测血瘀质肺结节良恶性表现最优的机器学习模型(area under curve (AUC)=0.893, 95% CI: 0.866~0.925). *Granulicatella*、*Fusobacterium*、*Delftia*、*Streptococcus*、*Alloprevotella*、*Veillonella*可作为血瘀质肺结节良恶性诊断的潜在生物标志物. 此外, 在血瘀质恶性肺结节患者中, 唾液微生物群可能存在有趣的代谢重编程. 因此, 唾液微生物可作为血瘀质恶性肺结节的非侵入性、高灵敏度鉴别诊断生物标志物, 为我国肺癌的临床诊疗提供一种更加精准、个体化的新方法.

关键词 肺结节, 液体生物标志物, 体质, 口腔微生物, 16S rRNA测序, 机器学习

针对高风险人群进行低剂量计算机断层扫描(low-dose computed tomography, LDCT)筛查是早期发现肺癌的主要手段, 及早检出肺结节显著降低了肺癌的死亡率^[1,2]. 然而, LDCT技术广泛使用的同时也面临着较高的假阳性率和过度诊疗的风险^[3], 导致部分患者接受了不必要的侵入性手术. 因此, 迫切需要针对恶性肺结节高风险人群开发非侵入性、高准确性的诊断

策略.

中医“体病相关”理论认为不同的体质类型与特定疾病的发生密切相关, 为肺结节高恶性风险人群识别提供了理论指导^[4,5]. 一项针对中国肺结节人群的中医体质调查显示, 血瘀质在肺结节人群中占有较大比例^[6]. 值得注意的是, 血瘀质被认为是多种肿瘤(包括肺癌)的高危体质, 且随年龄的增长, 血瘀质人群数量呈现

引用格式: 黄春霞, 马琼, 钟杨, 等. 基于唾液16S rRNA基因测序和机器学习算法确定血瘀体质肺结节良恶性诊断的微生物标志物. 科学通报, 2025, 70: 2002–2016

Huang C, Ma Q, Zhong Y, et al. Determine microbial markers for benign and malignant diagnosis of pulmonary nodules in blood stasis constitution based on saliva 16S rRNA gene sequencing and machine learning algorithms (in Chinese). Chin Sci Bull, 2025, 70: 2002–2016, doi: 10.1360/TB-2024-1314

逐渐上升趋势^[7]。这些证据表明,血瘀质可能是肺癌的高风险体质类型,在肺癌发生的早期阶段——肺结节,特别关注血瘀质人群具有重要意义。

近年来,生物标志物的发现和应用为肿瘤高风险人群的非侵入性评估提供了新的视角^[8]。特别是口腔微生物作为新兴的生物标志物,受到了研究者的广泛关注^[9]。一系列流行病学研究表明,口腔微生物正在成为影响肺癌发生发展的重要因素,与预后和疾病进展密切相关^[10,11]。更重要的是,多项研究表明口腔微生物群与中医体质类型之间具有较强的关联^[12],提示其可作为识别特定体质类型的重要指标。

尽管如此,关于血瘀质肺结节患者特异性口腔微生物标志物的研究仍然较为有限。本研究旨在通过16S rRNA基因测序技术,表征血瘀质良恶性肺结节患者唾液微生物群的差异,并结合机器学习(machine learning, ML)算法,从复杂的16S rRNA数据中识别一致且稳健的生物标志物。此外,我们初步揭示了唾液微生物与血瘀质肺结节之间的复杂关系,有助于为肺结节早期诊疗提供新的见解。

1 资料与方法

1.1 研究对象

受试者来自2022年6月~2024年6月期间在成都中医药大学附属医院、四川省肿瘤医院和成都市中西医结合医院的门诊患者。本研究遵循《赫尔辛基宣言》的原则,并获得了成都中医药大学附属医院伦理委员会的批准(伦理批号:2022KL-051)。该试验在中国临床试验注册中心进行了注册(注册号:ChiCTR2200062140;网址: <http://www.chictr.org.cn>)。

1.2 诊断标准

为了确保患者入组的科学性和准确性,每个中心由2名中医医师与3名西医医师(副主任医师以上)负责患者的中医体质评估、排除相关的基础疾病及临床病理与影像学诊断的确认。

(1) 肺结节诊断标准。参照《肺结节诊治中国专家共识(2024年版)》^[13],肺结节诊断标准为:影像学表现为直径 ≤ 3 cm的局灶性、类圆形、密度增高的实性或亚实性肺部阴影,可为孤立性或多发性,不伴肺不张、肺门淋巴结肿大和胸腔积液。

(2) 良恶性肺结节分类标准。通过活检或肺结节术

后病理结果判断结节的良恶性,以及2年内的影像学结果中结节消退或没有生长的证据来判断其为良性结节。良性恶性肺结节的病理结果判断参考2021年世界卫生组织(WHO)的《胸部肿瘤分类》^[14]。

(3) 中医体质分类与判定标准。参照2009年国家中医药管理局主持、中华中医药学会制定颁布的《中医体质分类与判定表》,本研究使用的体质判定标准为:患者根据量表中的所有问题作出相应回答,按照5级标准评分,每题根据各条目分会之和计算出原始分,再按照公式 $[(\text{原始分}-\text{条目数})/(\text{条目数}\times 4)]\times 100$ 计算出转化分,兼夹体质采用临界值判定方法^[15]。为了确保多中心入组患者血瘀体质判定的一致性,所有参与研究的中医医师统一经过了《中医体质分类与判定》标准的深入学习以及对具体辨证方法的规范操作培训,确保各中心使用相同的标准和方法对患者进行血瘀体质的评估。

1.3 纳排标准

纳入标准:(1) 所有患者均为首次就诊,并进行了详细的临床评估和影像学检查,符合肺结节西医诊断标准;(2) 符合中医血瘀体质的判定标准;(3) 自愿接受问卷调查,并经书面知情同意;(4) 年龄18~80岁。

排除标准:(1) 呼吸道感染、口腔疾病和其他并发症;(2) 有严重慢性基础疾病(如心脑血管疾病、糖尿病等);(3) 研究开始前一个月口服抗生素治疗;(4) 长期使用免疫抑制剂治疗;(5) 前3天摄入酸奶。

1.4 信息采集

一般信息采集:包括受试者的性别、年龄、吸烟史和癌症家族史等。

唾液样本采集:参与者在采样前至少2 h禁止受试者饮水和进食,用饮用水漱口,采样人员将唾液采集管(Salivette® Hygienic saliva collection)放在参与者的下唇上,让唾液自然流入管中。收集2~3 mL无刺激性唾液,在干冰中保存,并转移到 -80°C 的实验室冰箱中等待后续实验。

1.5 研究方法

1.5.1 微生物16S rRNA基因测序

唾液样本在干冰条件下送往上海美吉生物医药科技有限公司进行16S rRNA基因测序。使用EZNA® Soil Kit(Omega, Bio-tek, Norcross GA, USA)提取细菌DNA,

并通过1%琼脂糖凝胶电泳检测DNA质量,使用NanoDrop 2000C(Thermo Scientific, 美国)对DNA浓度和纯度进行评估. 使用上游引物338F(5'-ACTCC-TACGGGAGGCAGCAG-3')和下游引物806R(5'-GGACTACHVGGGGTWTCTAAT-3')对V3-V4 可变区的基因进行PCR扩增. 每个样本重复3次,建立基于扩增子的PE基因文库. 利用Illumina MiSeq PE250技术进行测序分析. 将MiSeq测序生成的PE reads根据重叠区域进行拼接,并通过fastp(v0.19.6)和flash(v1.2.7)进行数据的质控和拼接,获得优化后的数据. 随后,使用DADA2进行序列降噪,进行序列质量控制和过滤,从而获得扩增子序列变体(amplicon sequence variants, ASV)代表序列及其丰度信息. 物种注释采用Silva138/16s_bacteria数据库作为参考序列,并利用QIIME2中的classify_sklearn(naive bayes)算法对ASV进行物种注释,设置分类置信度为70%. 根据ASV分析结果,进一步进行多样性分析、群落构成分析以及群落间显著性差异分析等,并基于分类学信息在各个分类水平上进行群落结构统计.

1.5.2 微生物信息分析

利用Python 2.7.10统计两组样本中共有与独有的物种数目,并绘制Venn图;使用Mothur软件(v.1.30.2)计算不同随机抽样下的Alpha多样性指数(ace、shannon),并通过Wilcoxon秩和检验比较两组间的显著性差异. 使用R语言(v.3.3.1)绘制曲线图,基于Bray-Curtis距离进行β多样性差异分析,并分别进行主坐标分析(principal co-ordinates analysis, PCoA)和非度量多维尺度分析(non-metric multidimensional scaling, NMDS). 通过Wilcoxon检验筛选显著性物种,并利用LEfSe(http://huttenhower.sph.harvard.edu/galaxy/root?tool_id=lefse_upload)工具进行基于genus分类学组成的线性判别分析(linear discriminant analysis, LDA),设定LDA阈值为2. 基于Bray-Curtis距离,采用冗余分析(distance-based redundancy analysis, db-RDA)评估临床特征对唾液微生物群落的影响,并使用Spearman相关系数分析微生物群落与临床特征的相关性,绘制相关性热图.

菌群健康指数,是基于微生物组样本的物种水平分类特征来评估健康状况的稳健指数,侧重确定患病的可能性,独立于临床诊断;这是通过比较与良好和不良健康状况相关的两组微生物物种的相对丰度来实现的. 计算公式^[16]:

$$h_{i,M_H,M_N} = \log_{10} \left(\frac{\frac{R_{M_H}}{|M_H|} \sum_{j \in I_{M_H}} |n_j \ln(n_j)|}{\frac{R_{M_N}}{|M_N|} \sum_{j \in I_{M_N}} |n_j \ln(n_j)|} \right). \quad (1)$$

菌群失调指数,是确定微生物生态失调的程度指数,值越大说明菌群紊乱的程度越大. 计算公式: MDI = $-\log_{10}[(\text{total abundance in genera increased in disease group})/(\text{total abundance in genera decreased in disease group})]$.

1.5.3 模型构建与评价

采用了7种机器学习算法构建血瘀质良恶性肺结节预测模型.

(1) 逻辑回归(logistic regression, LR). 使用L2正则化(“penalty=‘l2’”)以防止过拟合,并设置了收敛容忍度(“tol=0.0001”)和最大迭代次数(“max_iter=100”). “C=1.0”表示正则化强度的倒数,“solver=‘lbfgs’”用于优化问题.

(2) 支持向量机(support vector machine, SVM). 选择径向基函数(“kernel=‘rbf’”)作为核函数,并设置了正则化参数“C=1.0”. “gamma=‘auto’”表示核函数的系数为特征数的倒数.

(3) 朴素贝叶斯(naïve bayes, NB). 使用伯努利朴素贝叶斯模型,设置了平滑参数“alpha=1.0”,并保留了先验概率(“fit_prior=True”).

(4) 随机森林(random forest, RF). 构建了100棵决策树(“n_estimators=100”),并设置了最小分割样本数“min_samples_split=2”.

(5) 梯度提升决策树(gradient boosting decision tree, GBDT). 设置学习率“learning_rate=0.1”,最大树深度“max_depth=5”,并构建了100棵决策树(“n_estimators=100”).

(6) 多层感知机(multi-layer perceptron, MLP). MLP模型由三层隐藏层组成,输入层维度为“embedd_dims”(即特征嵌入维度). 第一隐藏层包含128个神经元,第二隐藏层包含64个神经元. 每层之间通过线性变换(“nn.Linear”)连接. 采用LeakyReLU(“nn.LeakyReLU”)作为激活函数,以避免ReLU的“神经元死亡”问题,同时引入非线性能力. 使用Adam优化器进行训练,学习率设置为0.001. 损失函数根据任务需求选择二元交叉熵损失. 训练轮数通过五折交叉验证确定,以避免过拟合.

(7) 轻量梯度提升机(light gradient boosting machine, LightGBM). 设置最大树深度“max_depth=7”, 学习率“learning_rate=0.05”, 并构建了100棵决策树(“n_estimators=100”). “num_leaves=2 ** 5-1”表示每棵树的叶子节点数.

采用随机分层抽样的方法, 确保训练集和测试集中的正负样本比例是一致的, 从而将数据集划分为训练集(80%)和测试集(20%). 训练集用于模型的训练和参数优化, 而测试集用于评估模型的泛化性能. 为了进一步优化模型的参数, 我们还采用了五折交叉验证策略. 此外, 为了确保实验结果的稳定性和置信度, 使用不同的随机种子重复训练了30轮, 并取30轮的平均值作为最终结果. 通过绘制受试者特征曲线(receiver operating characteristic curve, ROC), 并利用曲线下面积(area under curve, AUC)评估各模型的诊断效能.

1.5.4 重要特征选择

将ML模型的特征重要性排序获得的微生物与SHapley Additive exPlanations (SHAP)筛选出的重要特征进行比较, 寻找出重要性特征一致的微生物.

SHAP基于博弈论中的Shapley值, 能够为每个特征分配一个重要性分数, 从而帮助我们理解模型的特征贡献. 具体步骤如下: (1) 模型选择: 基于最有预测效能的LightGBM模型进行SHAP分析; (2) SHAP值计算: 使用“shap.TreeExplainer (model)”对训练好的LightGBM模型进行解释, 并通过“shap_values=explainer.shap_values(X_test)”计算测试集(“X_test”)的SHAP值; (3) 特征重要性排序: 每个特征的SHAP值, 用于解释特征对模型预测的贡献度, 根据SHAP值的大小对特征进行排序, 筛选出对模型预测贡献最大的特征.

LightGBM通过计算每个特征在所有树中的加权分裂次数来衡量特征重要性. 特征在所有树中分裂节点的次数越多, 且分裂带来的损失函数减少越大, 重要性越高. 本研究将这3种不同的特征重要度排序作为初步的特征重要度参考依据, 用于对最终SHAP特征重要度的检验.

1.5.5 功能预测

在PICRUST2软件上基于京都基因与基因组百科全书(kyoto encyclopedia of genes and genomes, KEGG)和蛋白相邻类的聚簇(clusters of orthologous groups, COG)进行功能通路分析, 并使用MinPath确定基因通路. 首先将ASVs序列与其内部的参考序列比对, 把

ASVs置入相应地参考树中, 推断每个ASV基因家族的拷贝数, 预测每个ASV的基因含量, 确定每个样本基因家族的丰度. 然后, 将基因家族信息与对应的功能数据库(KEGG、COG)进行比对, 获得每个样本中对应的功能信息和丰度信息.

1.5.6 外部数据下载

我们从美国国立生物技术信息中心(National Center for Biotechnology Information, NCBI)下载了4个独立的外部验证数据集(<https://www.ncbi.nlm.nih.gov/>), 包括肺癌患者和健康受试者的唾液样本, 具体的Bioproject入库号为: 肺癌患者样本PRJNA904049、PRJEB29934、PRJNA687361, 健康受试者样本PRJNA1164710.

1.5.7 统计方法

采用Wilcoxon秩和检验、独立样本 t 检验、卡方检验和Fisher精确检验分别比较微生物组丰富度、多样性和临床特征方面的差异(SPSS® Statistics v22).

2 结果

2.1 患者队列和微生物组分析

经过LDCT报告和血瘀体质筛选, 纳入受试者331人. 使用16S rRNA基因测序对受试者的唾液样本进行了精确的ASV分析, 经过严格的质量控制和过滤, 最终确定了209个样本, 其中血瘀质良性肺结节(benign pulmonary nodules, BPN)患者136个, 恶性肺结节(malignant pulmonary nodules, MPN)患者73个(图1). 在受试者人口统计学和临床特征的比较方面(表1), 年龄、性别、吸烟史等基线特征方面, 两组间差异均无统计学意义($P>0.05$).

2.2 血瘀质良性和恶性肺结节唾液微生物的多样性改变

对微生物16S rRNA 基因序列进行稀释曲线分析(图2(a)), 结果显示每个样本的香农多样性指数(Shannon)的稀疏曲线达到了平台期, 表明大部分多样性已经获得. 总的来说, 鉴定了20个门、43个纲、93个目、175个科和382个属. 使用Shannon和Ace指数评估群落的 α 多样性(图2(b, c)), 血瘀质BMP组的 α 多样性显著高于BMP组. 对于 β 多样性, 基于主坐标分析(principal coordinates analysis, PCoA)和非度量多维尺度分析(non-metric multidimensional scaling, NMDS)进行分析(图2(d, e)). 结果显示, 血瘀质BPN组和MPN组之间的

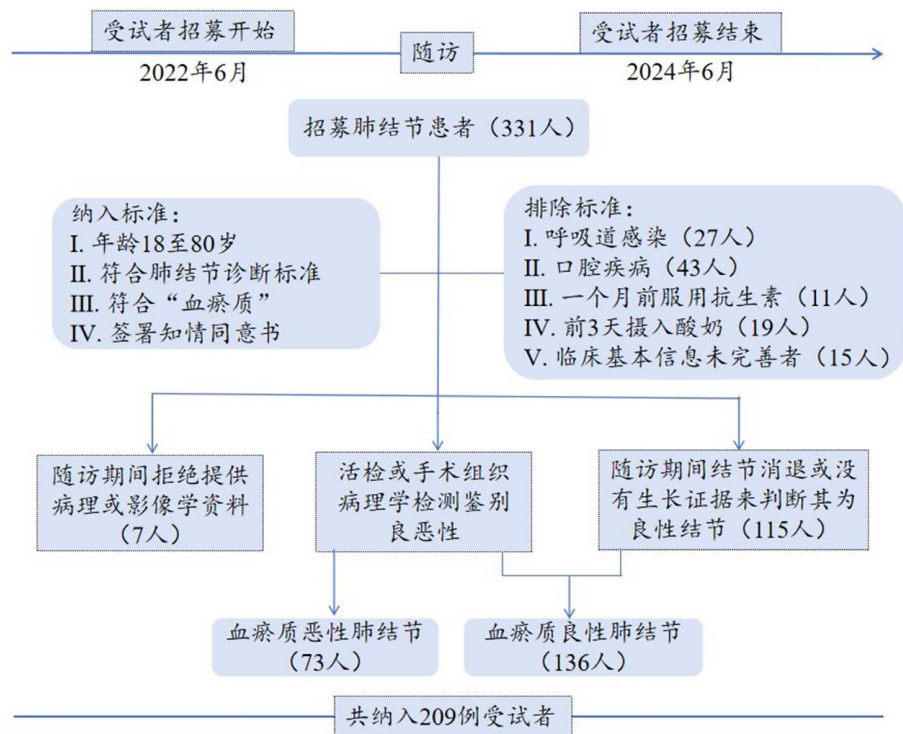


图 1 招募流程图
Figure 1 Recruitment flow chart

表 1 血瘀质BPN组和血瘀质MPN组的临床基本资料

Table 1 General characteristics of blood-stasis constitution group with BPN and MPN

临床特征	血瘀质BPN组($n=136$)	血瘀质MPN组($n=73$)	P
性别, n (%)			0.9840
女性	71(52.2)	38(52.1)	
男性	65(47.8)	35(47.9)	
年龄(岁, mean $\pm$ SD) ^{a)}	53.76 $\pm$ 9.15	54.95 $\pm$ 9.24	0.8933
吸烟史(%) ^{b)}			0.4714
仍然吸烟	16.04	10.87	
已戒烟	9.36	10.87	
从不吸烟	74.60	78.26	
个人肿瘤史(%) ^{c)}			<0.05
无	87.50	0	
有	12.50	100.00	

a) 秩和检验; b) 卡方独立性检验; c) Yates'卡方校正检验

微生物组显著分开, 且血瘀质MPN组的 β 多样性均低于BPN组(图2(f)). 同时, 利用rank-abundance curve分析观察了微生物多样性的丰度和均匀度(图2(g)), BPN组拥有更高的物种丰度和物种均匀度, 而更高的物种丰度和物种均匀度, 往往对宿主健康产生积极影响. 菌群健康指数(microbiome health index, MHI)(图2(h))和菌群

失调指数(microbiome dysbiosis index, MDI)(图2(i))的分析进一步支持了上述发现.

2.3 血瘀质良性肺结节的唾液微生物组成改变

进一步地, 我们还鉴定了血瘀质BPN和MPN人群唾液微生物群落的组成特征. 维恩图结果表明(图3(a)),

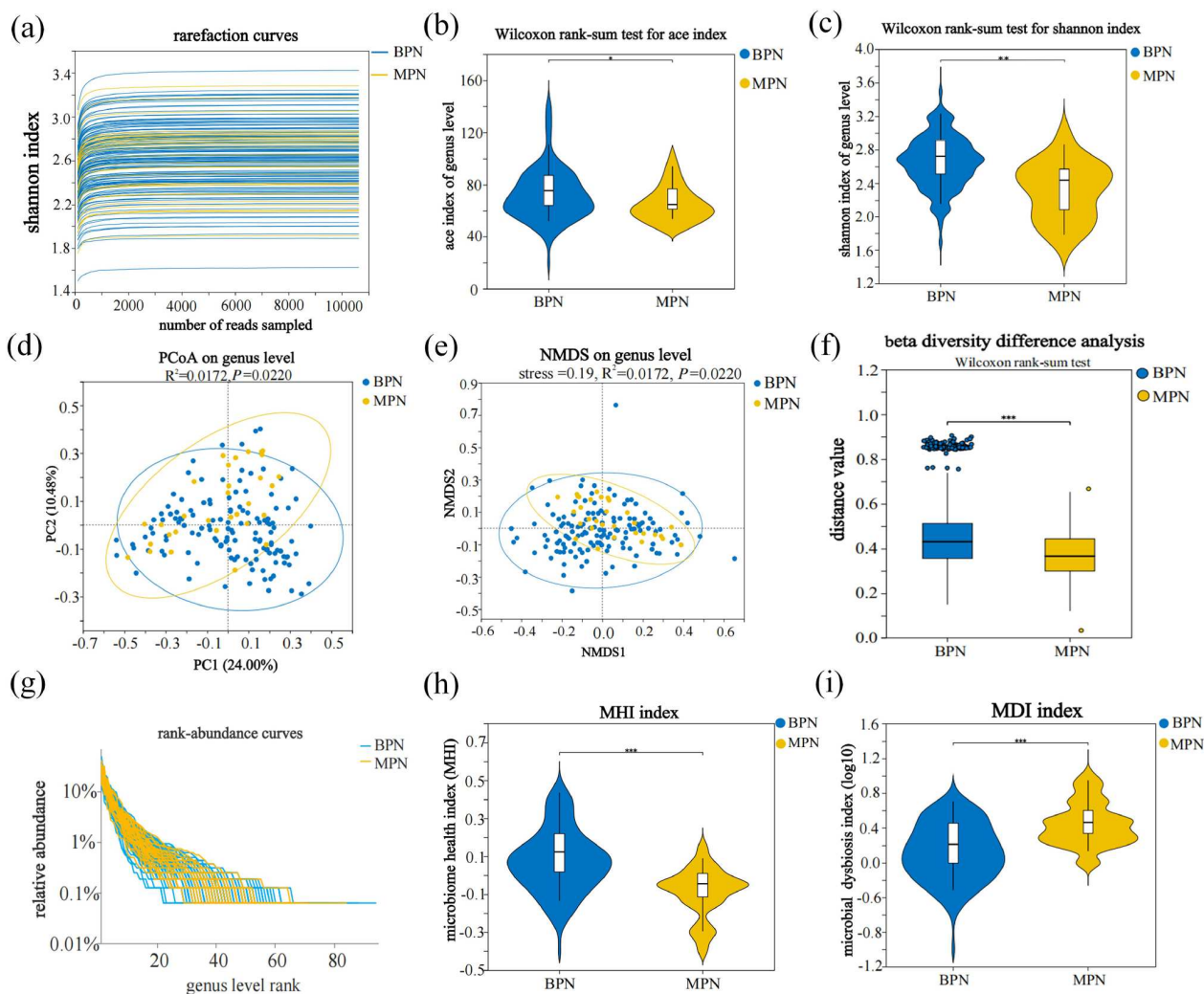


图 2 血瘀质BPN与MPN组的唾液微生物群 α 和 β 多样性. (a) 基于Shannon指数的稀释曲线分析曲线图. (b, c) 基于ACE指数(b)和Shannon指数(c)的血瘀质BPN和MPN组之间的 α 多样性比较. (d, e) 基于ADONIS/weighted-unifrac算法的NMDS分析($R^2=0.0172$, $P=0.0220$)对血瘀质BPN和MPN组的 β 多样性比较, 每个点代表一个样本, 具有相似微生物群组成的患者位于图的同一区域, 而彼此相距甚远的点代表具有不同微生物群的患者. (f) β 多样性组间差异检验. (g) 唾液样本的Rank-abundance分析曲线图. (h, i) 血瘀质BPN与MPN组唾液微生物群的健康指数(h)和失调指数(i)较高的健康指数值反映了菌群结构的稳定性和功能完整性, 与之相反, 较高的失调指数值通常与微生物群落结构失衡, 有益菌减少以及潜在病原菌增加相关联. BPN, benign pulmonary nodules; MPN, malignant pulmonary nodules

Figure 2 α and β diversity of salivary microbiota in BPN and MPN of blood stasis constitution. (a) Rarefaction curve analysis based on the Shannon index. (b, c) Comparison of α diversity between the BPN and MPN of blood stasis constitution based on the ACE index (b) and Shannon index (c). (d, e) β diversity comparison between the BPN and MPN of blood stasis constitution using NMDS analysis ($R^2=0.0172$, $P=0.0220$) based on the ADONIS/weighted-unifrac algorithm. Each point represents a sample, and patients with similar microbiota composition are located in the same area of the plot, while distant points represent patients with different microbiota. (f) β diversity inter-group difference test. (g) Rank-abundance curve analysis of saliva samples. (h, i) Health index (h) and dysbiosis index (i) of salivary microbiota in the BPN and MPN of the blood-stasis constitution. Higher health index values reflect the stability and functional integrity of the microbial community structure, whereas higher dysbiosis index values are typically associated with imbalances in the microbial community structure, reduced beneficial bacteria, and increased potential pathogenic bacteria. BPN: benign pulmonary nodules; MPN: malignant pulmonary nodules

血瘀质BPN和MPN患者均有其特异性口腔微生物组, 而血瘀质BPN患者具有更多的特有菌属数量. 随后, 基于属水平的相对丰度结果发现, 血瘀质MPN组与BPN组的优势菌群一致, 均为*Streptococcus*、*Rothia*、*Neis-*

seria, 但菌群占比有明显差异(图3(b)). 通过基于Wilcoxon秩和检验方法和基于LEfSe的判别分析方法对血瘀质BPN和MPN组之间的物种差异进行了双重检验验证, 结果显示了两组之间具有显著的差异微生物

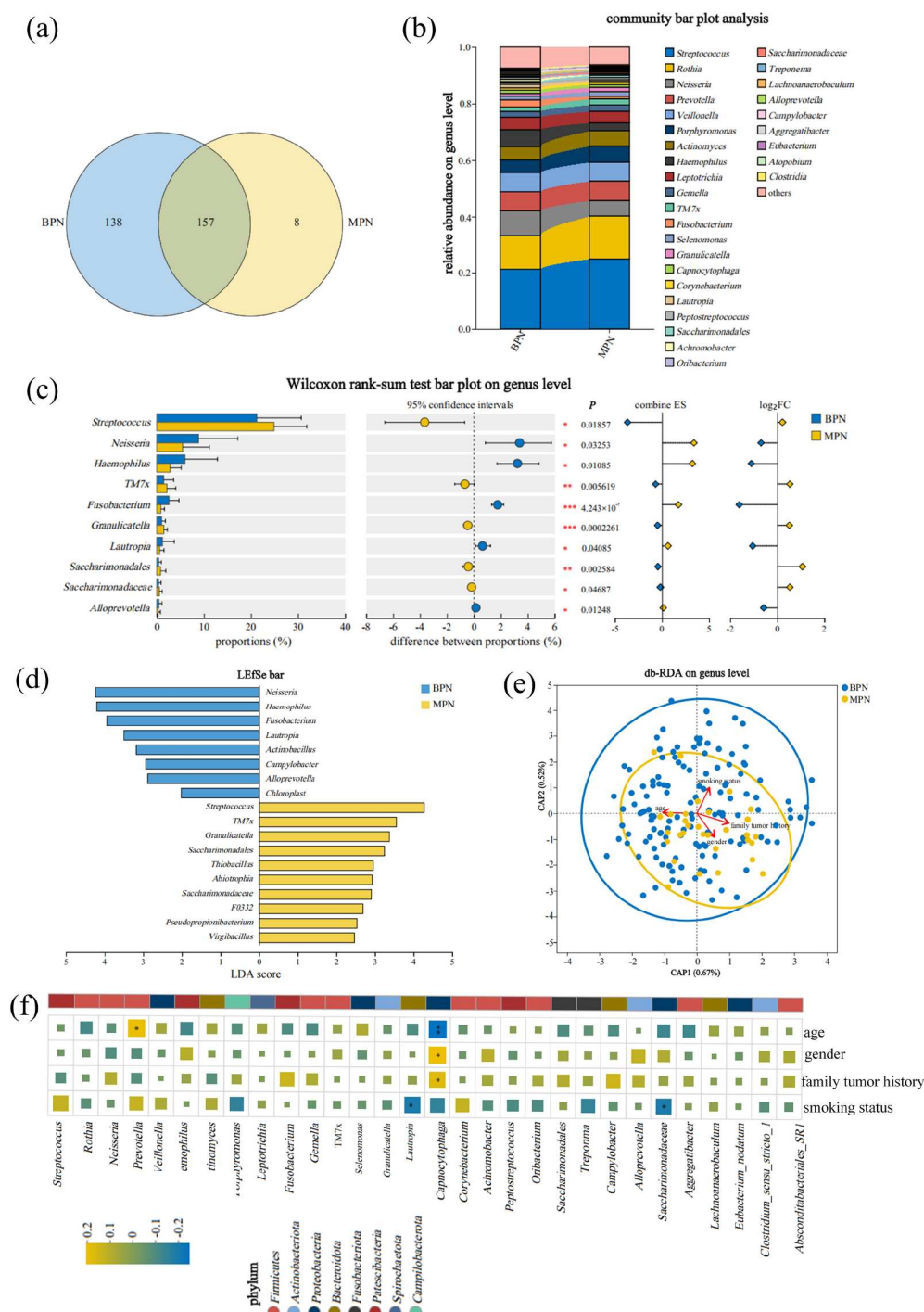


图 3 血瘀质良性肺结节患者的唾液微生物群主要成员的组成和分布. (a) 血瘀质BPN和MPN组之间唾液样本的菌属水平的维恩图. (b) 血瘀质BPN和MPN组唾液样本中属水平的微生物群结构组成. (c) 血瘀质BPN和MPN组唾液样本具有显著差异的菌群. (d) 通过LefSe分析生成的进化图显示, 两组唾液样本中类群丰富(LDA>2, $P<0.05$). (e) 基于Bray-Curtis距离的db-RDA分析. (f) 基于Spearman相关性分析的热图

Figure 3 Composition and distribution of the main members of the salivary microbiota in patients with benign and malignant pulmonary nodules of the blood-stasis constitution. (a) Venn diagram of the genus level of saliva samples between the BPN and MPN groups of blood stasis constitution. (b) The structural composition of the microbiota at the genus level in saliva samples from the BPN and MPN groups of blood stasis constitution. (c) Significantly different flora in saliva samples from the BPN and MPN groups of blood stasis constitution. (d) The evolutionary diagram generated by LefSe analysis shows that the two groups of saliva samples are rich in taxa (LDA score>2, $P<0.05$). (e) Db-RDA analysis based on Bray-Curtis distance. (f) Heatmap based on Spearman correlation analysis

群(图3(c, d)), *Streptococcus*、*TM7x*、*Granulicatella*、*Thiobacillus*、*Abiotrophia*、*Pseudopropionibacterium*、*Virgibacillus*等菌属在血瘀质MPN人群中丰度更高, 而 *Neisseria*、*Haemophilus*、*Fusobacterium*、*Lautropia*、*Actinobacillus*、*Campylobacter*、*Alloprevotella*等在血瘀质BPN人群中丰度更高。这些表明, 血瘀质BPN和MPN患者的唾液微生物组成改变, 可为区分提供较为丰富的微生物信息。

2.4 血瘀质良恶性肺结节唾液微生物组成与临床特征的相关性

我们进一步探讨了唾液微生物与血瘀质肺结节患者临床特征之间的关系。我们采用了距离基冗余分析(distance-based redundancy analysis, db-RDA)有效解析年龄、性别、吸烟史和家族肿瘤史等环境变量与微生物群落结构之间的关系。在本研究中, 这些临床特征对唾液微生物群落组成没有显著直接影响($P>0.05$) (图3(e)). 而进一步的Spearman相关性系数显示, 受试者临床特征与 *Capnocytophaga*、*Prevotella*、*Lautropia*等属具有显著相关性(图3(f)), 例如, *Prevotella*与吸烟状况呈现正相关, *Lautropia*与年龄呈现负相关, *Capnocytophaga*与性别、家族肿瘤史呈现正相关、而与吸烟状况呈现负相关。

2.5 基于唾液微生物组的血瘀质肺结节分类模型的构建和验证

由于每种ML方法在特征提取和模式识别方面的独特性, 我们对多种算法进行了评估, 以确定最适合我们唾液微生物数据集的模型。为此, 我们选择了7种具有代表性的ML分类器, 包含了经典且当前最先进的机器学习分类器: LR、SVM、MLP、NB、RF、GBDT、LightGBM。我们将209个唾液样本以8:2的比例随机分为两个队列, 队列一用于训练模型($n=167$), 队列二用于验证模型($n=42$)。这些队列在年龄、性别方面没有显示出统计学上的显著差异($P>0.05$), 确保了模型训练与验证的公平性。

我们观察了唾液微生物是否能区分血瘀质人群肺结节的良恶性表型, 结果显示所有模型的平均AUC为0.693~0.893(图4(a~g)), 表明基于唾液微生物组预测良恶性结节是可行的。为了评估预测模型的性能, 我们对7种模型之间的AUC值进行了比较(图4(h))。值得注意的是, 在测试集中LightGBM模型的性能显著优于所有

其他模型(AUC=0.893, 95% CI: 0.866~0.925)。

2.6 预测血瘀质肺结节良恶性的潜在唾液微生物标志物

在筛选血瘀质MPN分类中发挥关键作用的特定细菌类群, 我们采用了重要特征选择方法。通过运用最佳表现模型LightGBM进行重要特征排序, 我们成功识别出对肺结节良恶性诊断具有关键作用的6个微生物特征(图5(a)), *Granulicatella*、*Fusobacterium*、*Delftia*、*Streptococcus*、*Alloprevotella*、*Veillonella*, 这些特征与基于所有微生物特征的判别能力相当, 能够有效区分血瘀质BPN和MPN(图5(b))。接下来, 我们展示了6种潜在微生物标志物丰度在血瘀质BPN和MPN组的概率密度分布。核密度估计(kernel density estimate, KDE)曲线观察到, 6种标志物在BPN和MPN组中的分布存在差异。特别是, *Granulicatella*、*Streptococcus*、*Veillonella*在两组间的丰度峰值位置明显不同, *Delftia*、*Alloprevotella*、*Fusobacterium*在MPN组中呈现更尖锐的峰值, 表明在血瘀质MPN组中这些细菌的丰度更集中于特定范围(图5(c))。进一步地, 我们采用SHAP方法对每个关键微生物特征对模型预测的贡献度进行量化和解释。通过基于SHAP值大小的微生物特征排序, 我们发现这6个微生物特征在很大程度上与LightGBM重要性特征模式高度一致, 这进一步验证了它们在血瘀质肺结节良恶性分类中的关键作用(图5(d))。

在独立外部队列中评估筛选出的微生物标志物的预测效能, 可验证其准确性和稳定性。然而, 当前公共数据库尚无血瘀质肺结节人群的唾液微生物信息。因此, 我们从公共数据库中获取了健康受试者和肺癌患者的唾液微生物组数据, 包括102个肺癌样本和130个健康样本作为独立验证集, 进一步评估筛选出的微生物标志物的预测效能, 结果显示这些微生物标志物的菌群丰度在健康和肺癌患者之间表现出显著差异(图5(e))。进一步, 利用LightGBM模型对这些标志物进行了评估, 结果表明这些标志物在区分健康和肺癌疾病状态方面同样表现出较高的预测能力(AUC=0.871, 95%CI: 0.832~0.915), 这在一定程度上支持了其作为潜在诊断标志物的可行性(图5(f))。

2.7 口腔微生物组的功能概况

为了血瘀质MPN患者口腔微生物群的功能和代谢相关变化, 我们通过PICRUST2软件初步探索了血瘀质

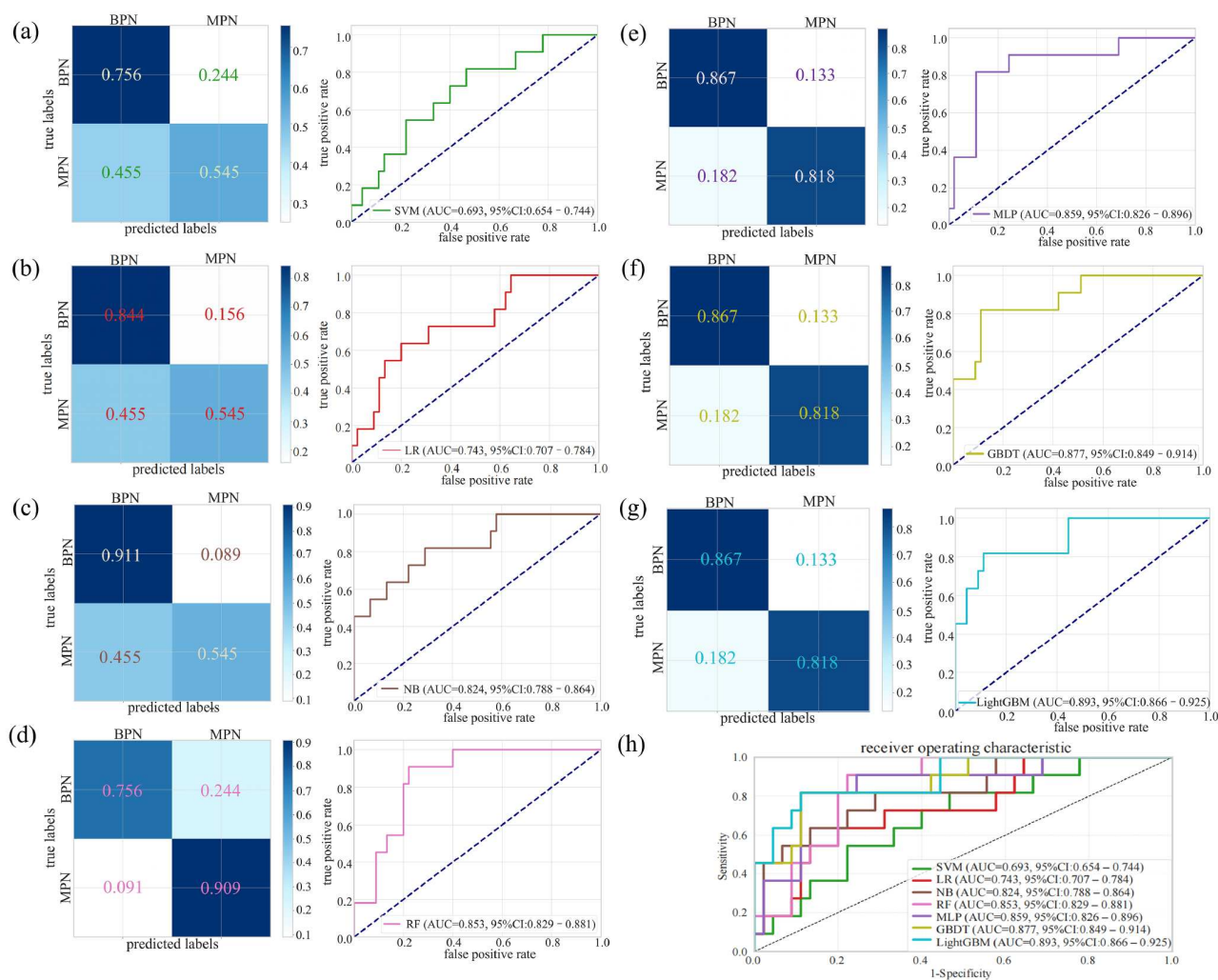


图4 基于唾液微生物信息的7种血瘀质肺结节分类模型的构建与评价。(a~g) 基于唾液微生物特征的SVM, LR, NB, RF, MLP, GBDT, LightGBM性能的归一化混淆矩阵和相应ROC曲线, 以评估模型对血瘀质BPN与MPN分类的准确性和预测效能, 归一化混淆矩阵展示了模型预测的真正例、假正例、真负例和假负例的比例, 用AUC提供了对模型整体性能的综合度量。(h) 7种预测模型的AUC曲线

Figure 4 Construction and evaluation of seven classification models for pulmonary nodules of blood stasis constitution based on salivary microbial information. (a~g) Normalized confusion matrices and corresponding ROC curves for the performance of SVM, LR, NB, RF, MLP, GBDT, and LightGBM based on salivary microbial features, to evaluate the accuracy and predictive efficiency of the models for classifying BPN and MPN of blood stasis constitution. The normalized confusion matrices display the proportions of true positives, false positives, true negatives, and false negatives predicted by the models, while AUC provides a comprehensive measure of the overall model performance. (h) AUC curves for the seven prediction models

BPN和MPN疾病状态下唾液微生物组改变情况。通过结合COG和KEGG数据库, 预测了在血瘀质BPN和MPN病变过程中代谢通路与直系同源类群的差异信息通路(图6(a~c)), 两组人群唾液微生物的功能组成高度相似, 但血瘀质MPN患者的唾液微生物群在氨基酸代谢、能量代谢和ABC型转运(ABC transporter)等途径中的丰度与血瘀质BPN组具有显著差别(图6(b, c))。这些数据表明, 在血瘀质肺结节的恶性发展过程中, 口腔

微生物群中可能存在有趣的代谢重编程。

3 讨论

不同体质具有特定的疾病易感性已是共识, 血瘀质作为包括肺癌在内的多种肿瘤的高危体质^[17,18], 在癌前病变的恶性进展中得到了更多的关注: 无论是在子宫内膜增生-子宫内膜不典型增生-子宫内膜癌的病程演进中, 体质偏向发生了痰湿质-湿热质-血瘀质的递

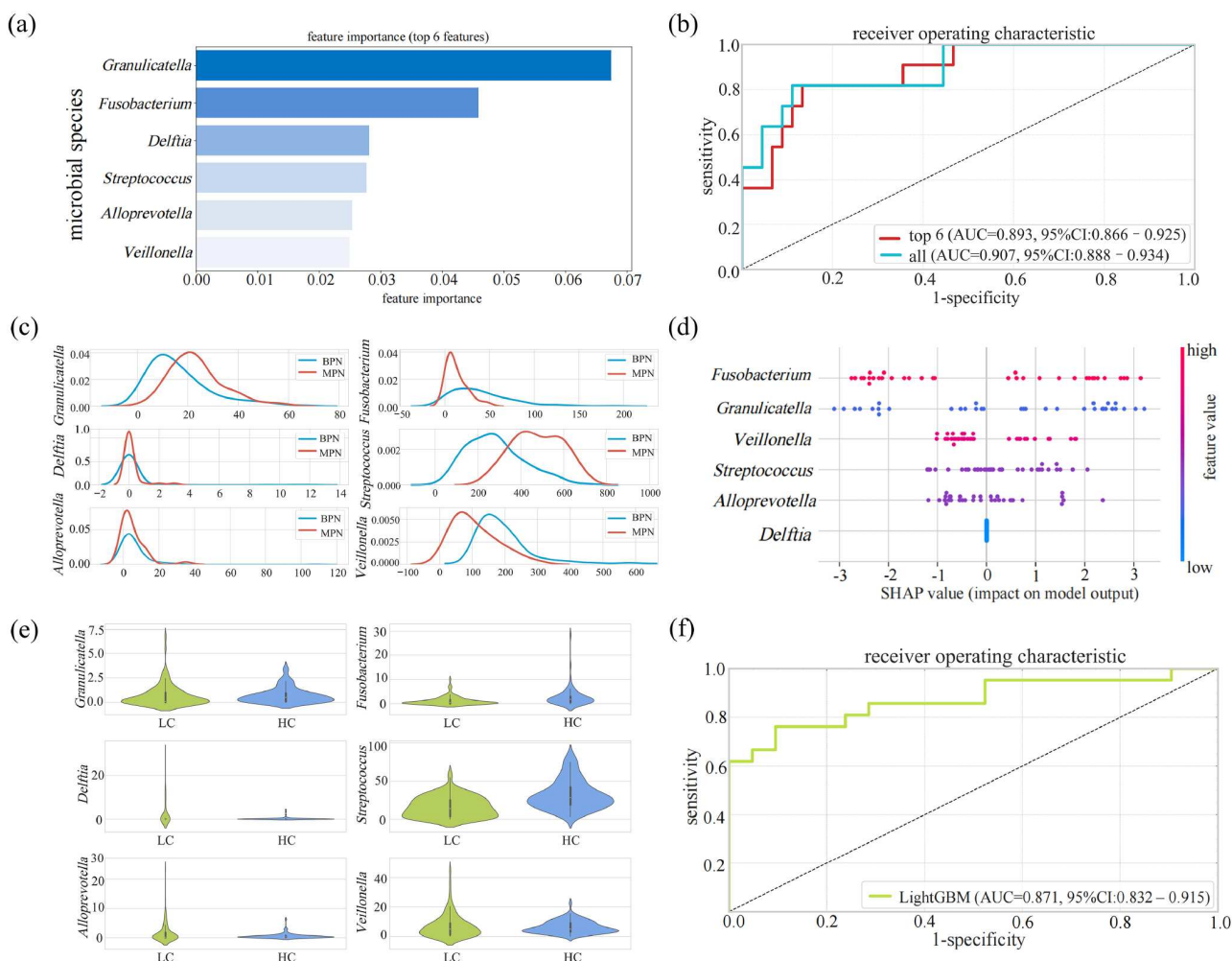


图 5 预测血瘀质肺结节良恶性的潜在唾液微生物标志物。 (a) 使用LightGBM模型的特征重要性绘制重要微生物的热图。 (b) 评估基于所有微生物特征与top6微生物特征的测试集ROC曲线。 (c) 血瘀质BPN和MPN组的KDE 曲线, 横轴代表细菌相对丰度, 纵轴代表概率密度, 曲线下面积为1, 两组曲线差异越大(例如峰值位置分离, 曲线形状不同), 表明该细菌丰度在两组间的差异越显著, 其作为区分良恶性肺结节标志物的潜力越大。 (d) 使用SHAP绘制两种样品类型中重要微生物的热图, SHAP值是由LightGBM模型的训练数据集进行测量的, SHAP值表示每个特征对模型预测的贡献程度, 正值表示增加预测值, 负值表示减少预测值, 根据SHAP值的大小对特征进行排序, 了解特征对预测的影响程度, 每一个数据点表示一组数据。 (e) 健康受试者与肺癌患者的菌群丰度差异图。 (f) 评估微生物标志物预测对健康受试者与肺癌患者区分效能的ROC曲线

Figure 5 Potential salivary microbial markers for predicting the BPN and MPN group of blood stasis constitution. (a) Heatmap of important microorganisms plotted using feature importance from the LightGBM model. (b) Evaluation of ROC curves for the test set based on all microbial features and the top6 microbial features. (c) KDE curves for BPN and MPN group of blood stasis constitution, with the x-axis representing bacterial relative abundance and the y-axis representing probability density. The area under the curve is 1. Greater differences between the two curves (such as separation of peak positions and different curve shapes) indicate more significant differences in bacterial abundance between the two groups, suggesting greater potential as a marker for distinguishing benign or malignant pulmonary nodules. (d) Heatmap of important microorganisms in two sample types plotted using SHAP values. SHAP values are measured from the training dataset of the LightGBM model and represent the degree of contribution of each feature to model predictions. Positive values increase the predicted value, while negative values decrease it. Features were sorted based on their SHAP values to understand their impact on predictions, with each data point representing a set of data. (e) Maps of differences in flora abundance between healthy subjects and lung cancer patients. (f) ROC curves were evaluated to predict the efficacy of microbial markers in distinguishing healthy subjects from lung cancer patients

演变化^[19]; 还是血瘀质患者更易发生肠化生, 进一步发展为胃癌概率更高^[20]; 抑或血瘀质患者相较于非血瘀质患者, 其肠道息肉更大、多数存在分叶、病理类型

多为腺瘤, 且复发率及癌变风险均更高^[21], 这些证据无不说明聚焦于肿瘤发生早期及演进过程中, 血瘀质人群无疑是更关键的因素。因此, 本研究以血瘀质肺结节

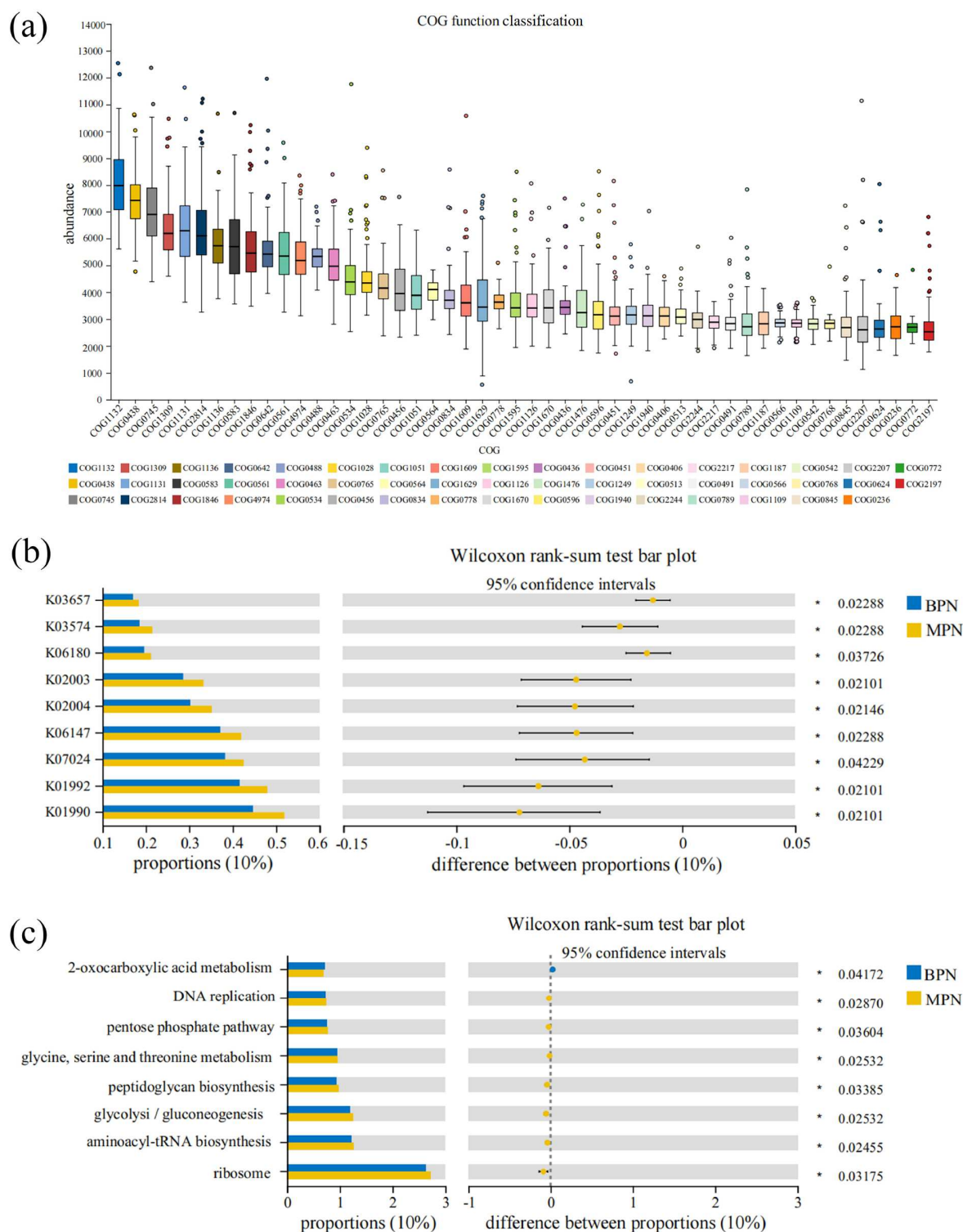


图 6 使用PICRUSt2预测口腔微生物群中的基因功能。(a) 基于COG数据库进行血瘀质BPN组和MPN组的基因差异功能分析。(b, c) 结合KEGG数据库预测血瘀质MPN组和BPN组的口腔微生物群基因功能差异(b), 并显示KEGG途径在三级(c)的结果

Figure 6 Using PICRUSt2 to predict gene functions in the oral microbiota. (a) Differential gene function analysis was performed for BPN and MPN group of blood stasis constitution based on the COG database. (b, c) Combining the KEGG database, the differences in gene functions of the oral microbiome between the BPN and MPN group of blood stasis constitution were predicted (b), and the results of the KEGG pathways at the third level were displayed (c)

患者作为研究对象,探索其非侵入性、高灵敏度的恶性诊断策略。

生物标志物的发现和应用为改善恶性肿瘤检测前概率的非侵入性估计提供了新的视角,并在多种肿瘤的早期良恶性诊断中得到了验证^[22,23]。据报道,口腔微生物群随着时间的推移是相当稳定的^[24]。多项研究证实,口腔微生物可作为疾病无创诊断的生物标志物,如口腔癌、结直肠癌和糖尿病等^[25-27]。因此,我们将受试者的唾液微生物组数据作为血瘀质良恶性肺结节的潜在生物标志物。 α 和 β 多样性能反应微生物的物种丰富度和相似性。许多研究都指出,较低的 α 和 β 多样性常与较高的肺癌风险相关^[10,28]。我们的研究与其结论一致,恶性肺结节的唾液微生物的 α 和 β 多样性指数显著降低。随后,微生物的组成成分结果同样证实了血瘀质BPN和MPN患者的唾液生物样本的组间具有较为丰富的差异菌属种类。据报道,更多的物种丰富度和丰度常被认为可为开发新的微生物标志物提供更广阔的研究空间。综合这些研究结果提示,唾液微生物特征具有生物标志物的潜力区分肺结节疾病状态。

微生物组数据具有维度高、不确定且过度分散等特征,这使传统统计工具难以利用微生物特征精确预测疾病。机器学习技术因其在处理复杂数据集方面的强大能力而逐渐被广泛应用于分析微生物群与疾病之间的关联^[29]。我们基于唾液微生物组数据,使用7种高度临床相关性且可重复的机器学习方法来评估血瘀质肺结节的良恶性风险。在模型性能的比较中,我们发现更先进的LightGBM在7个ML分类器中获得了最高的AUC值(0.893)。这与相关研究的结果一致,如Dong等基于简单的临床数据,利用多种机器学习算法构建了糖尿病肾病预测模型^[30],LightGBM在所测试的ML模型中具有最高的AUC值(0.815)。重要的是,相较于其他肺结节良恶性诊断模型^[31,32],我们的基于唾液微生物的LightGBM预测模型展现出更高的预测效能。

对疾病预测模型高贡献的微生物特征常被研究者们视为疾病诊断和未来机制研究的新生物标志物^[33]。我们利用最佳预测LightGBM模型识别出了最具区分性的微生物标志物,包括*Fusobacterium*、*Granulicatella*、*Veillonella*、*Streptococcus*、*Alloprevotella*、*Delftia*,而这6种微生物特征可以达到与所有特征一样的预测效能。据报道,这些微生物标志物与肺癌发生发展具有密切的联系。如Battaglia等人^[34]报道*Fusobacterium*在非小细胞肺癌患者中富集,与细胞毒性免疫反应的抑

制、干扰素 γ 基因表达的减少和MHC II类表达的减少有关,与ICB反应呈负相关。Hosgood等人^[35]对中国妇女口腔和痰标本进行了研究,发现肺癌患者与健康对照者相比,肺癌患者的*Granulicatella*和*Streptococcus*属明显富集。这与Zheng等人^[36]、Bello等人^[37]的研究一致,与良性肺结节相比,*Streptococcus*在肺癌患者中表现出更高的丰度。另一方面,SHAP算法提供了每种微生物对于分类器对样本的决策的重要性,可为每个参与者提供了更细粒度的信息。通过汇总参与者的SHAP值结果与LightGBM模型的特征重要性进行比较,结果显示这些微生物特征与模型具有相似的特征重要性。值得注意的是,使用独立验证集可以评估微生物标志物-模型的泛化能力,然而,目前尚无公开可用的相关数据。基于肺癌作为肺结节恶性演进的结局,我们从公共数据库中获取肺癌患者的唾液微生物数据作为验证集,结果显示这些标志物在区分健康和肺癌疾病状态方面同样表现出较高的预测效能。这些结果进一步表明,这些微生物特征与肺癌的发生和发展之间存在密切的关联。

此外,通过KEGG和COG数据库的PICRUSt2分析,我们初步探讨了恶性肺结节的潜在复杂机制。前沿证据表明,口腔致病微生物及其代谢产物能够异位定植到肺内,进而引起肺功能受损并促发肺部疾病^[38]。我们的分析发现,恶性肺结节中与能量代谢和氨基酸代谢直接相关的蛋白/分子功能富集最为明显,其丰度呈现出了不同程度的改变,COG功能分类的统计结果也支持了这一发现。众所周知,能量代谢在癌症发病机制中起着重要作用,且癌症系统中的代谢改变是促癌的关键因素之一^[39]。与此同时,我们还观察到恶性肺结节患者中氨基酸代谢途径的显著改变。已有研究表明这一改变可能通过mTOR、MYC、KRAS等信号通路调控表观遗传修饰,进而影响DNA和组蛋白的修饰,推动肿瘤进程^[40]。另外,ABC型转运系统信号通路丰度的增加进一步表明,口腔微生物群可能通过改变代谢过程在恶性肺结节中发挥作用。这一发现与肺癌作为代谢性疾病的假设相一致,肺癌患者常伴有代谢紊乱的表型^[41]。总体而言,上述实验结果与本研究的生物信息学分析结果相一致,支持了口腔微生物群在恶性肺结节发生中的潜在作用。

尽管我们的生物标志物组合和预测模型在评估血瘀质肺结节的恶性风险方面显示出良好的效果,但我们认识到研究存在一些局限性。首先,为了增强模型的

普适性和准确性,未来的研究应当扩展到更广泛的人群,纳入更多的临床变量,如年龄、吸烟史、肿瘤亚型、疾病阶段和种族等,这些因素在本研究中尚未得到充分考虑。其次,当前研究受限于样本量不足,特别是恶性肺结节的数据,以及对训练数据集的过度依赖。最后,口腔微生物群与宿主之间的复杂相互作用尚未得到充分验证,实验验证数据的缺乏限制了我们对这些相互作用的深入理解。未来,我们计划开展前瞻性临

床试验,以确保预测模型的稳定性和临床应用的准确性,并通过体内外实验验证微生物标志物的可靠性。

4 结论

本研究基于唾液微生物群构建的LightGBM模型,实现了对血瘀质肺结节的良恶性鉴别。这一创新尝试为开发针对血瘀质恶性肺结节的非侵入性、高灵敏度鉴别诊断生物标志物提供了新的可能。

参考文献

- Pastorino U, Silva M, Sestini S, et al. Prolonged lung cancer screening reduced 10-year mortality in the MILD trial: new confirmation of lung cancer screening efficacy. *Ann Oncol*, 2019, 30: 1162–1169
- Oudkerk M, Liu S Y, Heuvelmans M A, et al. Lung cancer LDCT screening and mortality reduction — evidence, pitfalls and future perspectives. *Nat Rev Clin Oncol*, 2021, 18: 135–151
- National Lung Screening Trial Research Team: aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*, 2011, 365: 395–409
- Lu W J, Zhou X M. Application of TCM constitution in early diagnosis and treatment of pulmonary nodules (in Chinese). *Shanxi J Tradit Chin Med*, 2018, 39: 518–520 [陆王娟, 周贤梅. 中医体质在肺结节早期诊治中的应用探讨. *陕西中医*, 2018, 39: 518–520]
- Liu H L, Yin G, Wang K, et al. Analysis of risk factors and TCM constitution types of malignant subsolid pulmonary nodules (in Chinese). *Hunan J Tradit Chin Med*, 2022, 38: 105–108 [刘浩雷, 尹格, 王克, 等. 恶性亚实性肺结节危险因素及中医体质类型分析. *湖南中医杂志*, 2022, 38: 105–108]
- Deng X J. The distribution and correlation of TCM constitution types and syndromes of pulmonary nodules (in Chinese). Master Thesis. Chang Chun: Changchun University Traditional Chinese Medicine, 2020 [邓小静. 肺结节中医体质类型和证素的分布规律及相关性研究. 硕士学位论文. 长春: 长春中医药大学, 2023]
- Chen S Q, Zi Y, Wei W, et al. The correlation between blood stasis constitution and diseases: a bibliometric analysis of 135 clinical studies (in Chinese). *J Tradit Chin Med*, 2024, 65: 529–536 [陈顺琪, 资艳, 魏巍, 等. 血瘀质与疾病的相关性研究——基于135项临床研究的文献计量分析. *中医杂志*, 2024, 65: 529–536]
- Hao L, Zhao R T, Welch N L, et al. CRISPR-Cas-amplified urinary biomarkers for multiplexed and portable cancer diagnostics. *Nat Nanotechnol*, 2023, 18: 798–807
- Kunath B J, De Rudder C, Laczny C C, et al. The oral–gut microbiome axis in health and disease. *Nat Rev Microbiol*, 2024, 22: 791–805
- Vogtmann E, Hua X, Yu G, et al. The oral microbiome and lung cancer risk: an analysis of 3 prospective cohort studies. *JNCI-J Natl Cancer Institute*, 2022, 114: 1501–1510
- Ma Q, Zeng X, Li X, et al. Current status and prospects of the relationship between oral microbiome and the occurrence and development of lung cancer (in Chinese). *Chin Sci Bull*, 2024, 69: 381–393 [马琼, 曾潇, 李雪珂, 等. 口腔微生物组与肺癌发生、发展的研究现状与展望. *科学通报*, 2024, 69: 381–393]
- Dong J. Study on the constitution of pulmonary nodules and the distribution characteristics of saliva microflora based on random forest algorithm (in Chinese). Master Thesis. Chengdu: Chengdu University of Traditional Chinese Medicine, 2023 [董敬. 基于随机森林算法的肺结节中医体质与唾液微生物群分布特征研究. 硕士学位论文. 成都: 成都中医药大学, 2023]
- Chinese Thoracic Society, Chinese Medical Association, Chinese Alliance Against Lung Cancer Expert Group. Chinese expert consensus on diagnosis and treatment of pulmonary nodules (2024) (in Chinese). *Chin J Tuberc Respir Dis*, 2024, 47: 716–729 [中华医学会呼吸病学分会, 中国肺癌防治联盟专家组. 肺结节诊治中国专家共识(2024年版). *中华结核和呼吸杂志*, 2024, 47: 716–729]
- WHO Classification of Tumours Editorial Board. WHO classification of tumours. Thoracic Tumours. 5th ed. Lyon: IARC Press, 2021
- Zhu Y B. Constitutional classification and comprehensive evaluation in Chinese medicine (in Chinese). *China J Tradit Chin Med*, 2012, 27: 40–42 [朱燕波. 中医体质分类判定与兼夹体质的综合评价. *中华中医药杂志*, 2012, 27: 40–42]
- Gupta V K, Kim M, Bakshi U, et al. A predictive index for health status using species-level gut microbiome profiling. *Nat Commun*, 2020, 11: 4635
- Cao Y J, Li Z P, Liu J P. The effect of COMT gene polymorphism and blood stasis constitution biological basis on the occurrence and development

- of lung cancer in mice (in Chinese). *J New Chin Med*, 2020, 52: 1–4 [曹月娇, 李志鹏, 刘加萍. COMT基因多态性及血瘀质生物学基础对小鼠肺癌发生发展的影响研究. *新中医*, 2020, 52: 1–4]
- 18 Zheng X, Meng X Z. The intervention effect of Feiyioma ointment on hemorheology of lewis tumor bearing mice with blood stasis (in Chinese). *Jiangxi Univ Tradit Chin Med*, 2015, 46: 39–40 [郑心, 孟宪志. 肺抑瘤膏对血瘀体质Lewis荷瘤小鼠血流变学的干预作用. *江西中医药*, 2015, 46: 39–40]
- 19 Li J Q. Study on the correlation between traditional Chinese medicine constitution and SIRT3 in endometrial disease (in Chinese). Master Thesis. Beijing: Beijing University Traditional Chinese Medicine, 2020 [李家琦. 子宫内膜病变的中医体质与SIRT3的相关性研究. 硕士学位论文. 北京: 北京中医药大学, 2020]
- 20 Xue Z F. Study on the correlation between blood congestion constitution in traditional Chinese medicine and neo immunohistochemical typing of gastric cancer and precancerous lesions (in Chinese). Master Thesis. Lanzhou: Gansu University Traditional Chinese Medicine, 2008. [薛增福. 中医淤血体质与胃癌、癌前病变新免疫组化分型的中基因CDX2、EGFR相关性研究. 硕士学位论文. 兰州: 甘肃中医学院, 2008]
- 21 Yang W G, Liao J L, Zhang Z J, et al. Pathological types of polyps and their prognosis in patients with colorectal polyps with blood stasis (in Chinese). *J Med Inf*, 2020, 33: 93–95 [杨文革, 廖金玲, 张志杰, 等. 瘀血质大肠息肉患者息肉病理类型及其预后研究. *医学信息*, 2020, 33: 93–95]
- 22 Zhou Y, Tao L, Qiu J, et al. Tumor biomarkers for diagnosis, prognosis and targeted therapy. *Sig Transduct Target Ther*, 2024, 9: 132
- 23 Chen X, Huang Y, Wee L, et al. Integrated analysis of radiomics, RNA, and clinicopathologic phenotype reveals biological basis of prognostic risk stratification in colorectal cancer. *Sci Bull*, 2024, 69: 3666–3671
- 24 Vogtmann E, Hua X, Zhou L, et al. Temporal variability of oral microbiota over 10 months and the implications for future epidemiologic studies. *Cancer Epidemiol Biomarkers Prevention*, 2018, 27: 594–600
- 25 Cai L, Zhu H, Mou Q, et al. Integrative analysis reveals associations between oral microbiota dysbiosis and host genetic and epigenetic aberrations in oral cavity squamous cell carcinoma. *npj Biofilms Microbiomes*, 2024, 10: 39
- 26 Liu Y, Wong C C, Ding Y, et al. *Peptostreptococcus anaerobius* mediates anti-PD1 therapy resistance and exacerbates colorectal cancer via myeloid-derived suppressor cells in mice. *Nat Microbiol*, 2024, 9: 1467–1482
- 27 Long J, Cai Q, Steinwandel M, et al. Association of oral microbiome with type 2 diabetes risk. *J Periodontal Res*, 2017, 52: 636–643
- 28 Hosgood H D, Cai Q, Hua X, et al. Variation in oral microbiome is associated with future risk of lung cancer among never-smokers. *Thorax*, 2021, 76: 256–263
- 29 Emmenegger B, Massoni J, Pestalozzi C M, et al. Identifying microbiota community patterns important for plant protection using synthetic communities and machine learning. *Nat Commun*, 2023, 14: 7983
- 30 Dong Z, Wang Q, Ke Y, et al. Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records. *J Transl Med*, 2022, 20: 143
- 31 Li L, Ye Z, Yang S, et al. Diagnosis of pulmonary nodules by DNA methylation analysis in bronchoalveolar lavage fluids. *Clin Epigenet*, 2021, 13: 185
- 32 Zheng Y, Dong J, Yang X, et al. Benign-malignant classification of pulmonary nodules by low-dose spiral computerized tomography and clinical data with machine learning in opportunistic screening. *Cancer Med*, 2023, 12: 12050–12064
- 33 Li M, Liu J, Zhu J, et al. Performance of gut microbiome as an independent diagnostic tool for 20 diseases: cross-cohort validation of machine-learning classifiers. *Gut Microbes*, 2023, 15: 2205386
- 34 Battaglia T W, Mimpfen I L, Traets J J H, et al. A pan-cancer analysis of the microbiome in metastatic cancer. *Cell*, 2024, 187: 2324–2335.e19
- 35 Hosgood III H D, Sapkota A R, Rothman N, et al. The potential role of lung microbiota in lung cancer attributed to household coal burning exposures. *Environ Mol Mutagen*, 2014, 55: 643–651
- 36 Zheng X, Sun X, Liu Q, et al. The composition alteration of respiratory microbiota in lung cancer. *Cancer Invest*, 2020, 38: 158–168
- 37 Bello S, Vengoechea J J, Ponce-Alonso M, et al. Core microbiota in central lung cancer with streptococcal enrichment as a possible diagnostic marker. *Archivos de Bronconeumologia (Engl Ed)*, 2021, 57: 681–689
- 38 Ma Q, Li X, Jiang H, et al. Mechanisms underlying the effects, and clinical applications, of oral microbiota in lung cancer: Current challenges and prospects. *Crit Rev Microbiol*, 2024, 50: 631–652
- 39 De Martino M, Rathmell J C, Galluzzi L, et al. Author correction: cancer cell metabolism and antitumour immunity. *Nat Rev Immunol*, 2024, 24: 537
- 40 Chen J, Cui L, Lu S, et al. Amino acid metabolism in tumor biology and therapy. *Cell Death Dis*, 2024, 15: 42
- 41 Nie M, Yao K, Zhu X, et al. Evolutionary metabolic landscape from preneoplasia to invasive lung adenocarcinoma. *Nat Commun*, 2021, 12: 6479

Summary for “基于唾液 16S rRNA 基因测序和机器学习算法确定血瘀体质肺结节良恶性诊断的微生物标志物”

Determine microbial markers for benign and malignant diagnosis of pulmonary nodules in blood stasis constitution based on saliva 16S rRNA gene sequencing and machine learning algorithms

Chunxia Huang^{1†}, Qiong Ma^{1†}, Yang Zhong¹, Xiao Zeng¹, Jiawei He¹, Xueke Li^{1,2}, Qian Wang¹, Xi Fu^{1,3}, Yifeng Ren^{1,3} & Fengming You^{1,3*}

¹ TCM Regulating Metabolic Diseases Key Laboratory of Sichuan Province, Hospital of Chengdu University of Traditional Chinese Medicine, Chengdu 610075, China

² Oncology Teaching and Research Department, Chengdu University of Traditional Chinese Medicine, Chengdu 610075, China

³ Institute of Oncology, Chengdu University of Traditional Chinese Medicine, Chengdu 610075, China

† Equally contributed to this work

* Corresponding author, E-mail: yfmdoc@163.com

Current diagnostic methods for pulmonary nodules are not ideal in clinical practice, especially for distinguishing benign from malignant nodules in high-risk populations. Therefore, there is an urgent need for precise and non-invasive diagnostic strategies. The oral microbiota, as an emerging biomarker, has received widespread attention, and advanced machine-learning algorithms provide new ideas for the development of non-invasive microbial markers for early identification of pulmonary nodules. In this study, saliva samples and information on traditional Chinese medicine constitution were collected from patients with pulmonary nodules. Based on histopathological detection of nodule tissue and the results of a blood stasis constitution scoring table, patients with benign ($n=136$) and malignant ($n=73$) pulmonary nodules of blood stasis constitution were selected as the study subjects. Using 16S rRNA sequencing technology and bioinformatics analysis, we revealed the oral microbial diversity, microbiota health index, microbiota dysbiosis index, species composition, differential bacterial genera, and the correlation between microorganisms and clinical characteristics of the subjects in patients with benign and malignant pulmonary nodules of blood stasis constitution. Based on the amplicon sequence variant of the salivary microbiota, seven machine learning techniques [logistic regression, support vector machine, multi-layer perceptron, naïve bayes, random forest, gradient boosting decision trees, light gradient boosting machine (lightGBM)] were used to construct a diagnostic model for pulmonary nodules of blood stasis constitution, and the predictive performance metrics were evaluated to determine the optimal performance model. The feature importance ranking based on the optimal model and the SHAP method was used to explain the contribution of features and screen for potential diagnostic biomarkers of pulmonary nodules in blood stasis constitution. Finally, based on PICRUSt2 functional prediction analysis of the microbiota, the potential mechanism of the oral microbiota in the malignant progression of pulmonary nodules was explored. The study suggested that salivary microbiota diversity of benign and malignant pulmonary nodules in blood stasis constitution was significantly different, with significantly reduced α and β diversity values in malignant pulmonary nodules ($P<0.05$). Simultaneously, compared to benign pulmonary nodules in blood stasis constitution, the microbiota health index was significantly reduced, and the microbiota dysbiosis index was significantly increased in patients with malignant pulmonary nodules ($P<0.05$). In terms of species composition, we found rich microbiota composition and high levels of differential microbiota in the saliva of patients with benign and malignant pulmonary nodules of blood stasis constitution, respectively. Additionally, Distance-based Redundancy Analysis showed a correlation between salivary microbiota composition and clinical characteristics in benign and malignant pulmonary nodules of blood stasis constitution. The seven machine learning models we developed were able to distinguish the malignancy probability of pulmonary nodules in blood stasis constitution. When validated on the model test dataset, the average performance AUC was 0.693–0.893, with LightGBM achieving high accuracy [area under the curve (AUC)=0.893, 95% CI: 0.866–0.925]. Based on the importance ranking of optimal model features and the SHAP method to explain feature contributions, potential biomarkers for predicting the risk of malignancy in pulmonary nodules of blood stasis constitution were identified as *Granulicatella*, *Fusobacterium*, *Delftia*, *Streptococcus*, *Alloprevotella*, and *Veillonella*. Finally, based on PICRUSt2 functional prediction analysis of the microbiota, we found interesting metabolic reprogramming that may exist in the oral microbiota during the malignant development of pulmonary nodules. Overall, machine learning models trained based on salivary microbiota features have good predictive performance. The oral microbiota can serve as a potential diagnostic biomarker for the differential diagnosis of pulmonary nodules in blood stasis constitution. This discovery opens new avenues for the development of non-invasive, highly accurate identification tools for high-risk populations with pulmonary nodules and deserves further research and clinical validation.

pulmonary nodule, liquid biomarker, body constitution, oral microorganism, 16S rRNA sequencing, machine learning

doi: [10.1360/TB-2024-1314](https://doi.org/10.1360/TB-2024-1314)