疾病监测
DISEASE SURVEILANCE

基于泛基因组特征的机器学习模型预测肺炎克雷伯菌对美罗培南的表型耐药

王瑜昊 赵俊岭 黄佳 吴鑫淼 卢昕 阚飙 李臻鹏

Establishment of machine learning models based on pan-genome features for prediction of phenotypic resistance of *Klebsiella pneumoniae* to meropenem

Wang Yuhao Zhao Junling Huang Jia Wu Xinmiao Lu Xin Kan Biao Li Zhenpeng

在线阅读 View online: https://doi.org/10.3784/jbjc.202411080633

耐药监测

基于泛基因组特征的机器学习模型预测肺炎克雷伯菌对美罗培南的表型耐药

开放科学（OSID）

王瑜昊[1]，赵俊岭[1]，黄佳[2]，吴鑫淼[1,3]，卢昕[3]，阚飙[3]，李臻鹏[3]

摘要： **目的** 建立基于全基因组特征的肺炎克雷伯菌对美罗培南表型耐药的机器学习模型，发现潜在耐药相关基因。**方法** 从细菌和病毒生物信息学资源中心数据库及美国国家生物技术信息中心的抗菌素耐药性生物体国家数据库数据库收集同时有表型数据和全基因组数据的菌株。使用 RGI 6.0.3 软件分析菌株基因组携带的耐药基因，使用 PanTa 1.0.0 软件分析菌株的泛基因组，分别使用耐药基因和附属基因作为纳入特征，构建预测肺炎克雷伯菌对美罗培南耐药表型的 LightGBM、随机森林、logistic 回归模型，使用分层嵌套交叉验证对模型进行特征筛选、超参数优化及模型评估，得到最优模型，使用 Shapley 可加性解释算法对特征的贡献进行评估。**结果** 经质量控制，有 5 800 株基因组纳入模型，其中对美罗培南耐药和敏感的菌株分别有 2 171 和 3 629 株，这些菌株包含泛基因 258 333 个，耐药基因 436 个。基于耐药基因分别构建的 3 种模型中，随机森林的拟合效果最佳。模型筛选到 64 种耐药基因，其曲线下面积（AUC）值、平衡准确度、召回率、特异度、精确度和阴性预测值分别为 0.916、87.84%、81.66%、94.02%、89.09% 和 89.55%。基于泛基因组构建的 3 种模型中，拟合效果最好的是了 logistic 回归。经过筛选得到了 156 种耐药相关的候选基因，其中包括 27 个已证实的耐药基因和 129 个潜在耐药相关基因，该模型优于耐药基因构建的模型，其 AUC 值、平衡准确度、召回率、特异度、精确度和阴性预测值分别 0.943、89.48%、85.16%、93.79%、89.16% 和 91.36%。进一步使用 Shapley 可加性解释算法评估了特征基因对模型的贡献。模型发现的前 15 个贡献最大的基因中有 6 个为未被证实的潜在耐药相关基因，如 *yojI*、*mobA*、*xerC* 和 *ynfE*。所建立的预测模型已被封装为命令行软件 predMemRes（https://github.com/Wangyuhao66/predMemRes），该软件能够基于肺炎克雷伯菌基因组序列快速预测菌株对美罗培南的耐药表型。**结论** 机器学习模型可有效用于预测肺炎克雷伯菌对美罗培南表型耐药，并发现潜在耐药相关基因。

关键词： 肺炎克雷伯菌；机器学习；美罗培南；耐药

中图分类号： R211；TP181 **文献标志码：** A **文章编号：** 1003-9961（2025）05-0653-07

Establishment of machine learning models based on pan-genome features for prediction of phenotypic resistance of *Klebsiella pneumoniae* to meropenem

Wang Yuhao[1], *Zhao Junling*[1], *Huang Jia*[2], *Wu Xinmiao*[1,3], *Lu Xin*[3], *Kan Biao*[3], *Li Zhenpeng*[3]. 1. *School of Public Health*, *Xinjiang Medical University*, *Urumqi* 830011, *Xinjiang*, China; 2. *Pathogenic Microorganism Identification Institute*, *Xinjiang Uyghur Autonomous Region Center for Disease Control and Prevention*, *Urumqi* 830002, *Xinjiang*, China; 3. *National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases*, *National Institute for Communicable Disease Control and Prevention*, *Chinese Center for Disease Control and Prevention*, *Beijing* 102206, China

Corresponding authors: *Zhao Junling,* Email: 378568407@qq.com; *Li Zhenpeng,* Email: lizhenpeng@icdc.cn

Abstract: Objective To establish machine learning models based on whole-genome features for the prediction of phenotypic resistance of *Klebsiella pneumoniae* to meropenem and identify potential resistance-associated genes. **Methods** The *K. pneumoniae* strains with both phenotypic data and whole-genome sequencing data were collected from the Bacterial and Viral Bioinformatics Resource Center and the Antibiotic Resistance Organism Reference Genome Database of the National Center for Biotechnology Information. Software RGI 6.0.3 was used to analyze the resistance genes carried by the strains, and software PanTa 1.0.0 was used to analyze the pangenome of the strains. LightGBM, Random Forest, and logistic regression models were constructed by using resistance genes and accessory genes as input features to predict the resistance phenotype of *K. pneumoniae* to meropenem. Stratified nested cross-validation was used for feature selection, hyperparameter optimization, and model evaluation to obtain the optimal model. Shapley additive exPlanations (SHAP) algorithm was used to evaluate the contribution of features. **Results** After quality control, 5 800 genomes were included in the models, with 2 171 meropenem resistant strains and 3 629 meropenem sensitive strains. These strains contained 258 333 pangenes and 436 resistance genes.

基金项目：国家重点研发项目（No. 2022YFC2303900）

作者单位：1. 新疆医科大学公共卫生学院，新疆 乌鲁木齐 830011；2. 新疆维吾尔自治区疾病预防控制中心病原生物鉴定所，新疆 乌鲁木齐 830002；3. 中国疾病预防控制中心传染病预防控制所，传染病溯源预警与智能决策全国重点实验室，北京 102206

作者简介：王瑜昊，女，河南省平顶山市人，在读硕士，主要从事传染病生物信息学研究，Email：1186442920@qq.com

通信作者：赵俊岭，Tel：13669926662，Email：378568407@qq.com；李臻鹏，Tel：010-58900744，Email：lizhenpeng@icdc.cn

收稿日期：2024-11-08 网络出版日期：2025-03-10

In the three models based on resistance genes, Random Forest model showed the best fit. The model identified 64 resistance genes with an area under the curve (AUC) value of 0.92, balanced accuracy of 87.84%, recall rate of 81.66%, specificity of 94.02%, precision of 89.09%, and negative predictive value of 89.55%. In the three models based on the pangenome, logistic regression model performed best. After screening, 156 candidate resistance-associated genes were identified, including 27 confirmed resistance genes and 129 potential resistance-associated genes. This model outperformed the model based on resistance genes, with an AUC value of 0.94, balanced accuracy of 89.48%, recall rate of 85.16%, specificity of 93.79%, precision of 89.16%, and negative predictive value of 91.36%. SHAP algorithm was further used to evaluate the contribution of feature genes to the models. In the top 15 genes with the greatest contributions identified by the model, six were unconfirmed potential resistance-associated genes, such as *yojI, mobA, xerC,* and *ynfE*. The established predictive model has been encapsulated into a command-line software named predMemRes (https://github.com/Wangyuhao66/predMemRes), which can rapidly predict the resistance phenotype of *K. pneumoniae* strains to meropenem based on their genome sequences. **Conclusion**　Machine learning models can effectively predict phenotypic resistance of *K. pneumoniae* to meropenem and identify potential resistance-associated genes.

Key words:　*Klebsiella pneumoniae*; Machine learning; Meropenem; Drug resistance

This study was supported by the fund for National Key Research and Development Program of China (No. 2022YFC2303900)

肺炎克雷伯菌属肠杆菌科是一种广泛存在于自然环境中的革兰阴性菌，为常见的院内感染潜在致病因子之一，能引起呼吸、泌尿、血液等多个系统的感染性疾病[1]。近年来，由于在临床治疗肺炎克雷伯菌感染中大量使用碳青霉烯类抗生素，碳青霉烯类耐药肺炎克雷伯菌菌株分离率正逐步上升[2–3]。美罗培南属于碳青霉烯类，其抗菌谱广泛，是治疗临床多药耐药菌感染的重要药物[4–6]。根据中国抗微生物药物监测网络的数据，肺炎克雷伯菌对美罗培南的感染率从 2005 年的 3.00% 上升到 2021 年的 23.10%。快速准确地检测肺炎克雷伯菌美罗培南耐药表型及了解耐药机制，可以为防治美罗培南耐药的发生和传播制定有效策略。

美罗培南耐药机制极其复杂，涉及多种因素，包括碳青霉烯酶产生、外排泵表达增加、外膜孔蛋白改变或减少[7]。其中最主要的原因是获得产碳青霉烯酶，能够水解碳青霉烯类抗生素，从而失去抗菌活性。碳青霉烯酶主要包括 A、B、D 3 类，A 类酶如 KPC 型在肺炎克雷伯菌中较为常见；B 类酶为金属酶，如 NDM、IMP、VIM 等；D 类酶主要为 OXA 型，如 OXA-48 等。传统方法难以全面了解肺炎克雷伯菌美罗培南的耐药机制和鉴定其潜在基因。机器学习可利用算法对数据进行自动学习和特征提取，进而完成预测、分类和决策任务，可基于全基因组序列数据与抗性表型之间的关系建立模型，快速检测耐药表型，识别潜在基因和深入理解耐药机制，从而弥补传统方法的局限性[8]。输入特征是基于机器学习算法的关键步骤，抗生素耐药性预测研究中，已知耐药基因、保守基因和单核苷酸多态性通常作为抗生素耐药性研究的输入特征[9–11]。除了输入特征外，模型的可解释性是机器学习应用在抗生素耐药性中的主要挑战。通过解释模型和筛选与耐药表型高度相关的遗传特征，可更好地将机器学习应用于临床实践。

随着基因组测序技术的普及及抗菌药物敏感性测试技术的进步，基因组特征结合耐药表型预测耐药表型取得了积极进展。Pesesky 等[12] 在 2016 年率先使用机器学习及规则的方法基于肠杆菌科菌株的全基因组测序数据预测了菌株对 12 种抗生素的耐药表型，预测结果与表型一致性达到 90.00% 左右。Macesic 等[13] 在 2020 年使用 600 余株肺炎克雷伯菌 CG258 的基因组序列预测了其对多黏菌素的表型耐药，并识别到了多个已知的多黏菌素耐药基因和几个可能对模型有重要贡献的非耐药基因。Xu 等[14] 整合 3 928 株肺炎克雷伯菌的基因组及多种抗生素耐药表型数据构建 Lasso 回归预测其耐药表型，预测模型曲线下面积（area under the curve，AUC）大于 0.900，并在宏基因组测序数据上进行了评估应用。Pataki 等[15] 使用 704 个大肠埃希菌基因组，整合基因组突变及获得性耐药基因信息构建了预测大肠埃希菌对环丙沙星的最低抑菌浓度（minimum inhibitory concentration，MIC），并取得了较好的预测效果。

由于不同的物种和药物组合可能需要不同的机器学习模型来实现最佳的预测效果，本研究将选择 3 种经典的机器学习算法开展，包括 LightGBM[16]、随机森林和 logistic 回归。LightGBM 综合了包括 XGBoost 在内的梯度提升决策树算法框架内各算法的优势[17]。随机森林算法通过集成多个决策树的预测结果来提高整体的预测性能和鲁棒性[18]。logistic 回归是一种常用的解决分类问题的统计模型，给出事件发生的概率，用于预测事件发生的可能性。

本研究分别以肺炎克雷伯菌的耐药基因及泛基因组为特征建立 3 种机器学习模型，预测其对美罗培南耐药的表型，使用 SHapley 可加性解释算法

（SHapley additive exPlanations，SHAP）解释重要贡献的关键特征，以探索潜在的新型耐药基因，为肺炎克雷伯菌美罗培南耐药机制的研究和耐药特征的鉴定提供了新的思路。

1 材料与方法

1.1 材料 美罗培南耐药表型数据是从细菌和病毒生物信息学资源中心及抗菌素耐药性生物体国家数据库收集的具有抗生素药敏试验数据的全基因组。基因组数据是从美国国家生物技术信息中心数据库中下载肺炎克雷伯菌基因组序列。

1.2 方法

1.2.1 指标判定 MIC 阈值和断点依据临床实验室标准协会（Clinical and Laboratory Standards Institute，CLSI）标准。美罗培南的 MIC 值≤1 μg/mL 判定为敏感，＞4 μg/mL 判定为耐药。根据该标准鉴定美罗培南对肺炎克雷伯菌的表型。

1.2.2 质量控制 采用 Kleborate 3.0.0 软件进行质量控制[19]，通过准确识别物种和序列类型，确定关键的获得性遗传特征，最终筛选出 5 800 株鉴定为肺炎克雷伯菌的菌株。

1.2.3 泛基因组分析 用 PanTA 1.0.0 软件对 5 800 株肺炎克雷伯菌进行泛基因组分析[20]。PanTA 能够逐步构建泛基因组，无需从头开始重建已累积的集合，可根据基因在集合中的普遍程度，将得到的基因簇分类为核心基因或附属基因。

1.2.4 耐药基因预测 使用基于抗生素耐药性综合数据库（the comprehensive antibiotic resistance database，CARD）的 RGI 6.0.3 软件对耐药基因进行预测[21]，以识别编码抗生素抗性蛋白基因。

1.2.5 预测模型的构建及评估 为防止模型构建过程中的信息泄露使模型出现过拟合现象，本研究编写了 R 脚本结合使用 tidymodels 软件包进行模型构建，使用分层嵌套交叉验证的方法分别构建和评估 LightGBM、随机森林、logistic 回归模型，并对 3 个模型进行比较，从而选择最优模型。具体流程如下：外层将数据集分为 5 份，4 份做内层，1 份为外层。内层使用 3 折交叉验证，数据集分为 3 份，其中 2 份作为训练集，1 份作为验证集。内层数据用于特征选择、参数优化，外层用于模型评估。

（1）特征选择：模型构建过程中使用特征选择降低模型的复杂度，提高模型的泛化能力，同时减少计算成本。本研究使用多种特征选择的方法联合进行筛选特征：①对于耐药基因特征，采用 Fisher 精确性检验对耐药组和敏感组差异的耐药特征进行分析，并对 P 值进行本雅明尼–霍奇金（Benjamini-Hochberg，BH）矫正，保留 $P<0.01$ 的显著特征；②对于泛基因特征，使用 Fisher 精确性检验分析耐药组和敏感组差异泛基因，并对 P 值进行 BH 矫正，保留 $P<0.01$ 的显著特征；使用最小绝对收缩和选择算子回归（least absolute shrinkage and selection operator regression，LASSO）对泛基因筛选，进一步缩小特征的范围；③对于耐药基因特征和泛基因特征，进一步使用 RFE 算法逐步移除每次迭代中对模型贡献最小的特征，得到最佳特征集。

（2）超参数调优：使用格子搜索的方法进行超参数调优。logistic 回归模型超参数为 penalty，其取值大于 0；随机森林设定了 mtry、trees、min_n 3 个超参数，mtry 的为 2～50，trees 和 min_n 使用默认调优范围；LightGBM 设置的超参数为 mtry、tree_depth、learn_rate、min_n、loss_reduction。mtry 被限定在 2～6，其他参数使用默认调优范围。

（3）模型评估：LightGBM、随机森林、logistic 回归模型经过超参数优化后获得的最优模型，外层使用 5 折交叉验证进行模型评估，模型评估指标包括混淆矩阵，AUC 值，平衡准确度、特异度、精确度、F1 分数、阴性预测值、召回率等。其中，F1＝2×(精确度×召回率)/(精确度+召回率)

1.3 统计学分析 使用 SHAP 评估特征对模型的贡献，SHAP 值使用基于 R 语言的 fastshap 软件包计算得到。使用基于 R 语言的 shapviz 包绘制特征重要性蜂群图，显示不同特征对模型预测结果的总体影响。绘制特征重要性柱状图，展示每个特征在模型中的相对重要性。为评估分类模型的整体性能，直观显示模型在各类阈值下的表现，使用 ROCit 包绘制受试者工作特征（receiver operating characteristic，ROC）曲线，将正类样本从负类样本中区分出来，ROC 曲线越靠近左上角，说明模型的性能越好。

2 结果

2.1 样本描述 在纳入的 5 800 株肺炎克雷伯菌菌株中，地理和时间覆盖范围较为广泛。样本来源于全球 6 个大洲，包括亚洲（1 479 株）、欧洲（1 870 株）、北美洲（1 851 株）、大洋洲（49 株）、非洲（26 株）、南美洲（12 株），涉及 60 个国家；样本收集时间为 1981—2024 年，2010—2014 年收集的样本量最多，占总体的 60.20%；来源包括人类、动物和环境，主要的来源是人类，有 5 185 株，占总体的 89.40%。根据 CLSI 标准对美罗培南耐药的有 2 171 株，对美罗培南敏感的有 3 629 株。

2.2 基于耐药基因构建预测美罗培南表型的随机森林模型 使用基于 CARD 数据库的 RGI-6.0.3 软件和

resfinder 4.0 软件对耐药基因进行预测，分别得到了 436 和 405 个耐药基因，使用较全面的 RGI 的预测结果进行后续分析。本研究使用分层嵌套交叉验证的方法进行超参数优化和模型评估，外层的 5 折交叉验证用于模型评估，模型构建的流程图见图 1。对基于耐药基因构建的 LightGBM、随机森林、logistic 回归 3 种模型拟合效果进行比较，结果显示随机森林模型的拟合效果最好，AUC 值和 F1 分数均最高。使用递归特征消除算法（recursive feature elimination，REF）移除最小贡献特征，在随机森林模型中筛选出 64 个最优特征，作为数据集进行机器学习的模型构建。见表 1。

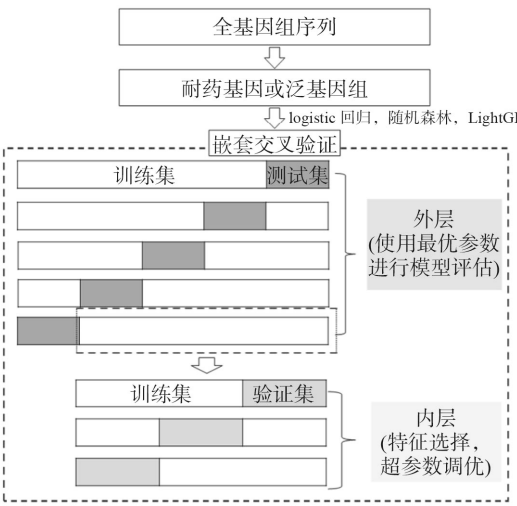


图 1　机器学习模型构建流程图

Figure 1　Flowchart for machine learning model construction

特征对于模型预测结果影响的分析显示，*KPC-2*，*KPC-3*，*OXA-48* 和 *NDM-1* 等特征对随机森林模型评估贡献较大。蜂群图中特征的散点较为分散，可能是区分耐药和敏感菌株的关键因素。不同特征对随机森林模型预测结果的总体影响不同，蜂群图的方向表示特征对预测结果的平均影响方向，负特征值表示该特征值增加时，模型更可能预测为耐药；正特征值则相反。如 *KPC-2*、*KPC-3*、*OXA-48* 和 *BRP*（*MBL*）等，当这些特征的值增加时，菌株的表型可能预测为耐药；而 *TEM-1*、*sul2* 等这些特征的值增加时，模型则更可能预测为敏感。见图 2。

随机森林模型混淆矩阵的灵敏度为 81.66%，特异度为 94.02%，精确度为 89.09%，阴性预测值为 89.55%，灵敏度和阳性预测值的调和平均数 F1 分数为 0.921，平衡准确度为 87.84%。以上指标表明该模型在总体上拟合效果良好。

ROC 曲线展示随机森林模型在不同阈值设置下的性能。ROC 曲线紧密贴近左上角，显示较高的分类性能，此 ROC 曲线的 AUC 为 0.916，远高于随机分类器的 AUC 值 0.500，表示模型具有出色的区分正负样本的能力。见图 3。

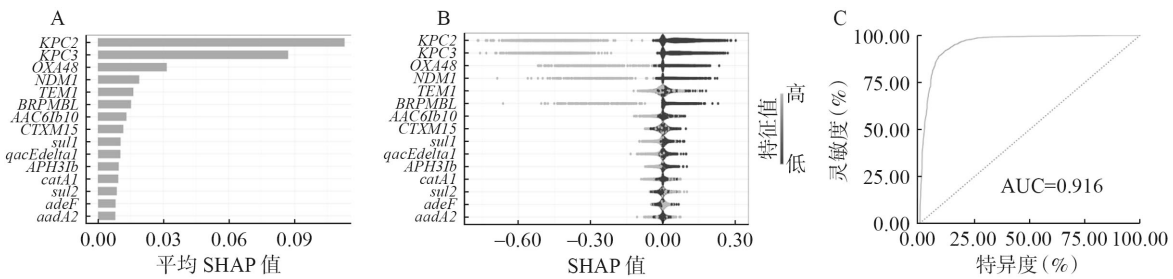
2.3　基于泛基因组构建预测美罗培南表型耐药的机器学习模型及比较　泛基因组分析得到 258 333 个同源基因，通过 BLAST 2.14.0 比对，筛选掉与 CARD 数据库预测序列比对上的基因，再和耐药特征组成新的输入特征，本研究使用分层嵌套交叉验证的方法进行超参数优化和模型评估，外层的 5 折交叉验证用于模型评估，LightGBM、随机森林、logistic 回归 3 种模型的拟合效果显示，logistic 回归模型表现最佳，灵敏度、阴性预测值、AUC 值和平衡准确度均最高。经过 RFE 算法筛选得到 156 个特征用于构建最终 logistic 回归模型。这 156 个特征中有 29 个是耐药基因，129 个是新发现的可能导致肺炎克雷伯菌对美罗培南表型耐药的潜在基因。见表 2。

重要性柱状图展示了特征对 logistic 回归模型

表 1　基于耐药基因构建的 3 种机器学习模型效果比较

Table 1　Comparison of effectiveness of three machine learning models based on drug resistance genes

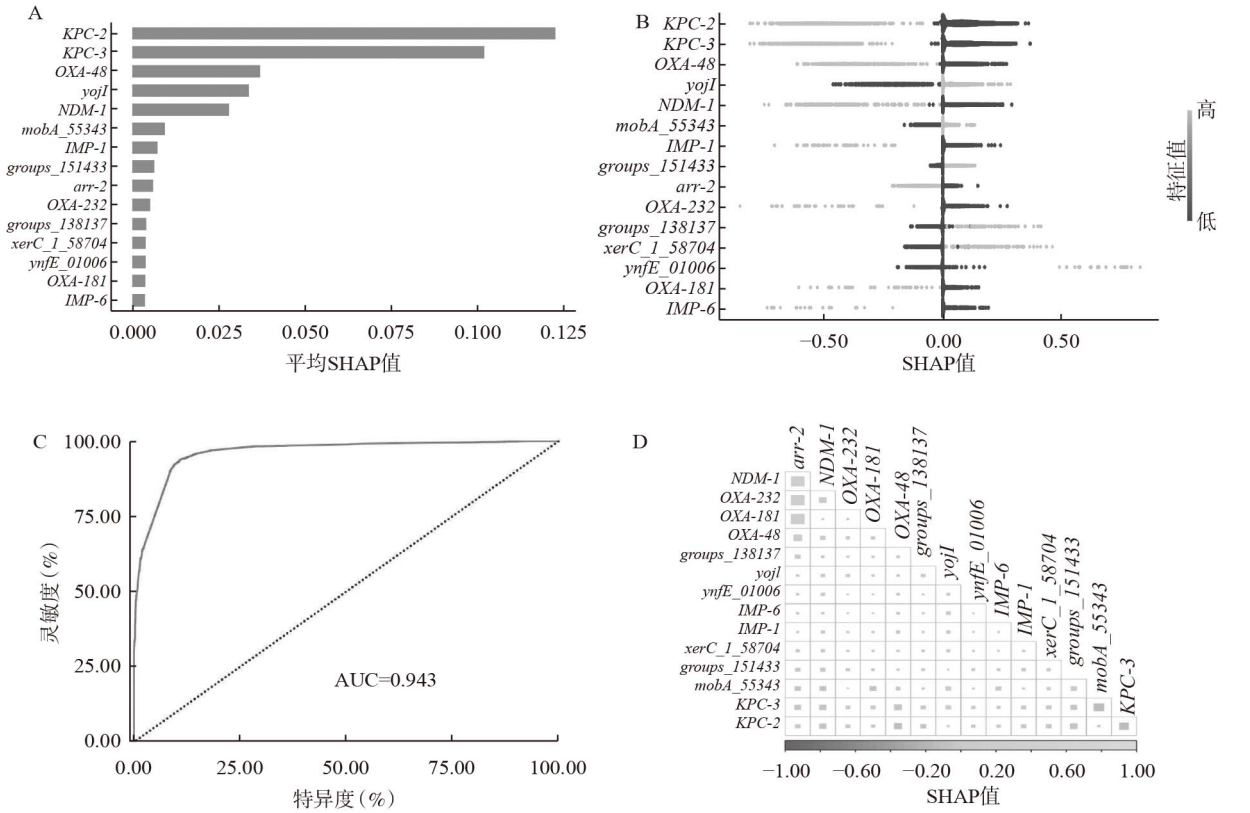
模型	灵敏度（%）	特异度（%）	精确度（%）	阴性预测值（%）	曲线下面积	F1分数	平衡准确度（%）
LightGBM	80.88	94.43	89.71	89.20	0.907	0.917	87.66
随机森林	81.66	94.02	89.09	89.55	0.916	0.921	87.84
logistic回归	82.31	93.99	89.13	89.88	0.903	0.892	88.15



注：A. 特征重要性柱状图；B. 特征重要性蜂群图；C. 模型受试者工作特征曲线图；SHAP. SHapley 可加性解释算法；AUC. 曲线下面积。

图 2　基于耐药基因的随机森林模型评估

Figure 2　Evaluation of Random Forest model based on drug resistance genes



注：A. 特征重要性柱状图；B. 特征重要性蜂群图；C. 模型受试者工作特征曲线图；D. SHAP 值排名前 15 位的基因相关矩阵图；SHAP. SHapley 可加性解释算法；AUC. 曲线下面积。

图 3　基于泛基因组的 logistic 回归模型评估

Figure 3　Evaluation of logistic regression model based on pan-genome data

表 2　基于泛基因组构建的 3 种机器学习模型效果比较

Table 2　Comparison of effectiveness of three machine learning models based on pan-genome

模型	灵敏度（%）	特异度（%）	精确度（%）	阴性预测值（%）	曲线下面积	F1分数	平衡准确度（%）
LightGBM	82.77	94.38	89.85	90.15	0.943	0.949	88.58
随机森林	83.78	93.69	88.83	90.62	0.936	0.938	88.73
logistic回归	85.16	93.79	89.16	91.36	0.943	0.944	89.48

预测结果的影响，结果显示 *KPC-2*、*KPC-3*、*OXA-48* 和 *yojI* 等特征对模型有较大贡献，在区分耐药和敏感菌株时起到了更大的作用。蜂群图点的位置反映了该特征对预测结果的贡献，颜色深浅表示特征值的大小，可颜色的变化反映特征值对预测结果的影响趋势，颜色越浅代表特征值越大，反之，则代表特征值越小。*KPC-2*、*KPC-3*、*NDM-1* 和 *IMP-1* 等特征增加了耐药性的预测概率；而 *ynfE_01006*、*groups_138137* 等特征降低了耐药性的预测概率。见图 3。

从 logistic 回归模型混淆矩阵的结果来看，灵敏度为 85.16%，特异度为 93.79%，精确度为 89.16%，阴性预测值为 91.36%，F1 分数为 0.94，平衡准确度为 89.48%，该模型在预测肺炎克雷伯菌对美罗培南的表型耐药方面具有较高的准确度、灵敏度和特异性，且一致性良好。以上指标表明模型在实际应用中可能具有较高的可靠性和实用性。

与使用 CARD 数据库的直接预测相比，基于泛基因组的分类模型其 ROC 曲线更加靠近左上角，几乎与坐标轴重合，AUC 值为 0.94，进一步证实了其出色的预测能力。对 SHAP 值较高的 15 个基因进行功能注释，除了已经明确的耐药基因外，主要涉及一些酶类，如 ABC 转运蛋白渗透酶、钼辅因子有关转移酶、酪氨酸重组整合酶及二甲亚砜还原酶亚基 A，还有 DNA 包装蛋白及假蛋白，这些基因的功能可能与耐药存在密切关联。见表 3。

根据每个基因在每个样本对模型贡献的 SHAP 值，对这 15 个基因两两之间进行相关性分析，方块的大小和颜色表示相关性的强度和符号。结果显示，两两之间有 29 对特征的 SHAP 值存在显著相关；其中有 14 对是已被证实为耐药基因之间的，有 13 对存在于已被证实的耐药基因与耐药相关基因之间，有 2 对属于耐药相关基因之间，这些基因之间对模型的贡献可能存在一定的交互作用。

表 3　基于泛基因组构建的逻辑回归模型 SHAP 值排名前 15 位的基因

Table 3　Top 15 genes ranked by SHAP values of logistic regression model based on pan-genome

基因	SHAP 评分	功能	是否为已被证实的耐药基因
KPC-2	0.123	肺炎克雷伯菌碳青霉烯酶	是
KPC-3	0.104	肺炎克雷伯菌碳青霉烯酶	是
OXA-48	0.038	D类碳青霉烯酶	是
yojI	0.034	ABC转运蛋白渗透酶	否
NDM-1	0.027	金属β-内酰胺酶	是
mobA_55343	0.009	钼辅因子有关转移酶	否
IMP-1	0.007	金属β-内酰胺酶	是
groups_151433	0.007	假蛋白	否
arr-2	0.006	编码β-抑制蛋白	是
OXA-232	0.005	β-内酰胺酶	是
OXA-181	0.004	D类碳青霉烯酶	是
groups_138137	0.004	DNA包装蛋白	否
xerC_1_58704	0.004	酪氨酸重组整合酶	否
ynfE_01006	0.004	二甲亚砜还原酶亚基A	否
IMP-6	0.004	金属β-内酰胺酶	是

注：SHAP. SHapley可加性解释算法；ABC转运蛋白.依赖腺嘌呤核苷三磷酸水解将物质由细胞内转移至细胞外的跨膜蛋白。

2.4　肺炎克雷伯菌对美罗培南耐药表型预测程序的开发　本研究所建立的预测模型已被封装成基于 linux 操作系统的命令行程序 predMemRes，并上传至 GitHub 网站（https://github.com/Wangyuhao66/predMemRes），该软件能够通过输入肺炎克雷伯菌基因组序列，快速预测菌株对美罗培南的耐药表型。预测结果包括 2 个文件，一个是表型预测（耐药/敏感）及对应概率值，另一文件列出检测到的耐药基因及相关耐药基因。使用方法如下：./predMemRes.sh -a genome.fasta -o output_tag，其中参数 a 为输入肺炎克雷伯菌的基因组序列（FASTA 格式），参数 o 为输出结果标签。

3　讨论

快速预测抗菌素耐药谱对大多数严重细菌感染的临床管理至关重要，同时对确定最佳抗生素治疗方案具有广泛意义。随着临床数据的逐渐增长和算法性能的提高，机器学习方法在抗生素耐药预测方面具有很大的应用潜力[22]。基因型–表型模型成功地以高灵敏度和特异度预测了结核分枝杆菌（MTB）[23-24]、大肠埃希菌[25]、肺炎克雷伯菌[26]、铜绿假单胞菌[27]和其他物种的耐药表型[28]。本研究考虑到肺炎克雷伯菌对美罗培南的耐药机制主要与基因获得相关，将耐药基因和泛基因有无作为特征构建模型，分别建立 lightGBM 模型、随机森林模型、logistic 回归模型，以探索潜在的耐药相关基因。

通过分析最优模型的最佳特征集发现，*yojI*、*xerC*、*ynfE* 等 129 个基因可能与肺炎克雷伯菌对美罗培南表型耐药相关。其中，*yojI* 作为外排泵，被证实在大肠埃希菌中对抗菌肽 J25 的耐药有关[29]，有研究发现，*xerC* 的缺失增加了金黄色葡萄球菌对多种抗生素的敏感性，以及对宿主免疫防御的敏感性[30]。有研究报道 *ynfE* 对 *MarA* 介导的多重耐药非常重要[31]。本研究中这些基因对模型有较高的贡献，提示其可能影响美罗培南的耐药性。

使用基于机器学习的方法预测病原细菌的表型耐药有多种优势：首先该方法可以基于现有基因组数据实现快速表型预测，尤其适用于难培养病原体的测序数据分析及对已积累大量测序数据的菌种可开展规模化耐药性筛查。其次，该方法可以发现不同于其他碳青霉烯类药物的补充耐药机制。

已有多项研究建立了肺炎克雷伯菌全基因组特征与亚胺培南[32]、多黏菌素[13]等抗生素表型耐药的关联模型，但本研究通过纳入更全面的数据集拓展了分析维度，但本研究也存在一定的局限性；当前模型仅整合耐药基因和泛基因组数据，未来纳入单核苷酸多态性、基因拷贝数变异等特征可能提升预测效能；异质性数据来源可能导致系统误差；鉴于肺炎克雷伯菌对美罗培南的耐药性处于动态进化中，需建立模型迭代更新机制以适应未来数据积累。

综上所述，机器学习算法在预测肺炎克雷伯菌对美罗培南耐药性方面所展现出的强大能力，为深入理解耐药机制提供了新的视角和方法，以及开发潜在的治疗策略奠定了基础信息。

利益冲突　所有作者均声明不存在利益冲突

参 考 文 献

[1] Iannaccone M, Boattini M, Bianco G, et al. Meropenem/vaborbactam-based combinations against KPC-producing *Klebsiella pneumoniae* and multidrug-resistant *Pseudomonas aeruginosa*[J]. *Int J Antimicrob Agents*, 2020, 56(2): 106066. DOI: 10.1016/j.ijantimicag.2020.106066.

[2] Podschun R, Ullmann U. *Klebsiella* spp. As nosocomial pathogens: Epidemiology, taxonomy, typing methods, and pathogenicity factors[J]. *Clin Microbiol Rev*, 1998, 11(4): 589−603. DOI: 10.1128/cmr.11.4.589.

[3] Satlin MJ, Chen L, Patel G, et al. Multicenter clinical and molecular epidemiological analysis of bacteremia due to carbapenem-resistant *Enterobacteriaceae*（CRE）in the CRE epicenter of the united states[J]. *Antimicrob Agents Chemother*, 2017, 61(4): e02349−16. DOI: 10.1128/aac.02349−16.

[4] Couffignal C, Pajot O, Laouénan C, et al. Population pharmacokinetics of imipenem in critically ill patients with suspected ventilator-associated *pneumonia* and evaluation of dosage regimens[J]. *Br J Clin Pharmacol*, 2014, 78(5): 1022−1034. DOI: 10.1111/bcp.12435.

[5] Chen HH, Xue YH, Shen WW, et al. Epidemiology and resistance mechanisms to imipenem in *Klebsiella pneumoniae*: a multicenter study[J]. *Mol Med Rep*, 2013, 7(1): 21−25. DOI: 10.3892/mmr.2012.1155.

[6] Chen LF, Anderson DJ, Paterson DL. Overview of the epidemiology and the threat of *Klebsiella pneumoniae* carbapenemases（KPC）resistance[J]. *Infect Drug Resist*, 2012, 5: 133−141. DOI: 10.2147/idr.S26613.

[7] Chen L, Mathema B, Chavda KD, et al. Carbapenemase-producing *Klebsiella pneumoniae*: Molecular and genetic decoding[J].

Trends Microbiol, 2014, 22(12): 686−696. DOI: 10.1016/j.tim.2014.09.003.
[8] Anahtar MN, Yang JH, Kanjilal S. Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research[J]. *J Clin Microbiol*, 2021, 59(7): e0126020. DOI: 10.1128/jcm.01260−20.
[9] Kuang XY, Wang F, Hernandez KM, et al. Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN[J]. *Sci Rep*, 2022, 12(1): 2427. DOI: 10.1038/s41598−022−06449−4.
[10] Aytan-Aktug D, Nguyen M, Clausen PTLC, et al. Predicting antimicrobial resistance using partial genome alignments[J]. *mSystems*, 2021, 6(3): e0018521. DOI: 10.1128/mSystems.00185−21.
[11] Nguyen M, Olson R, Shukla M, et al. Predicting antimicrobial resistance using conserved genes[J]. *PLoS Comput Biol*, 2020, 16(10): e1008319. DOI: 10.1371/journal.pcbi.1008319.
[12] Pesesky MW, Hussain T, Wallace M, et al. Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in gram-negative bacilli from whole genome sequence data[J]. *Front Microbiol*, 2016, 7: 1887. DOI: 10.3389/fmicb.2016.01887.
[13] Macesic N, Walk IV OJBD, Pe'er I, et al. Predicting phenotypic polymyxin resistance in *Klebsiella pneumoniae* through machine learning analysis of genomic data[J]. *mSystems*, 2020, 5(3): e00656−19. DOI: 10.1128/mSystems.00656−19.
[14] Xu YP, Liu DL, Han P, et al. Rapid inference of antibiotic resistance and susceptibility for *Klebsiella pneumoniae* by clinical shotgun metagenomic sequencing[J]. *Int J Antimicrob Agents*, 2024, 64(2): 107252. DOI: 10.1016/j.ijantimicag.2024.107252.
[15] Pataki BÁ, Matamoros S, van der Putten BCL, et al. Understanding and predicting ciprofloxacin minimum inhibitory concentration in *Escherichia coli* with machine learning[J]. *Sci Rep*, 2020, 10(1): 15026. DOI: 10.1038/s41598−020−71693−5.
[16] Hajihosseinlou M, Maghsoudi A, Ghezelbash R. A novel scheme for mapping of MVT-type Pb–Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm[J]. *Nat Resour Res*, 2023, 32(6): 2417−2438. DOI: 10.1007/s11053−023−10249−6.
[17] Wu W, Li MS, Wu Y, et al. Cluster energy prediction based on multiple strategy fusion whale optimization algorithm and light gradient boosting machine[J]. *BMC Chem*, 2024, 18(1): 24. DOI: 10.1186/s13065−024−01127−0.
[18] Jin YD, Lan AL, Dai YR, et al. Development and testing of a random forest-based machine learning model for predicting events among breast cancer patients with a poor response to neoadjuvant chemotherapy[J]. *Eur J Med Res*, 2023, 28(1): 394. DOI: 10.1186/s40001−023−01361−7.
[19] Lam MMC, Wick RR, Watts SC, et al. A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex[J]. *Nat Commun*, 2021, 12(1): 4188. DOI: 10.1038/s41467−021−24448−3.
[20] Le DQ, Nguyen TA, Nguyen SH, et al. Efficient inference of large prokaryotic pangenomes with PanTA[J]. *Genome Biol*, 2024, 25(1): 209. DOI: 10.1186/s13059−024−03362−z.
[21] Alcock BP, Raphenya AR, Lau TTY, et al. Card 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database[J]. *Nucleic Acids Res*, 2020, 48(D1): D517−D525. DOI: 10.1093/nar/gkz935.
[22] Wang SY, Zhao CJ, Yin YY, et al. A practical approach for predicting antimicrobial phenotype resistance in *Staphylococcus aureus* through machine learning analysis of genome data[J]. *Front Microbiol*, 2022, 13: 841289. DOI: 10.3389/fmicb.2022.841289.
[23] Walker DTM, Kohl TA, Omar SV, et al. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: A retrospective cohort study[J]. *Lancet infect Dis*, 2015, 15(10): 1193−1202. DOI: 10.1016/S1473−3099（15）00062−6.
[24] The CRyPTIC Consortium and the 100, 000 Genomes Project. Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing[J]. *N Engl J Med*, 2018, 379(15): 1403−1415. DOI: 10.1056/NEJMoa1800474.
[25] Humphries RM, Bragin E, Parkhill J, et al. Machine-learning model for prediction of cefepime susceptibility in *Escherichia coli* from whole-genome sequencing data[J]. *J Clin Microbiol*, 2023, 61(3): e01431−22. DOI: 10.1128/jcm.01431−22.
[26] Stoesser N, Batty EM, Eyre WD, et al. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data[J]. *J Antimicrob Chemother*, 2013, 68(10): 2234−2244. DOI: 10.1093/jac/dkt180.
[27] Kos VN, Déraspe M, McLaughlin RE, et al. The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility[J]. *Antimicrob Agents Chemother*, 2015, 59(1): 427−436. DOI: 10.1128/AAC.03954−14.
[28] Su M, Satola SW, Read TD. Genome-based prediction of bacterial antibiotic resistance[J]. *J Clin Microbiol*, 2019, 57(3): e01405−18. DOI: 10.1128/JCM.01405−18.
[29] Socías SB, Vincent PA, Salomón RA. The leucine-responsive regulatory protein, Lrp, modulates microcin J25 intrinsic resistance in *Escherichia coli* by regulating expression of the YojI microcin exporter[J]. *J Bacteriol*, 2009, 191(4): 1343−1348. DOI: 10.1128/jb.01074−08.
[30] Ledger EVK, Lau K, Tate EW, et al. XerC is required for the repair of antibiotic- and immune-mediated DNA damage in *Staphylococcus aureus*[J]. *Antimicrob Agents Chemother*, 2023, 67(3): e01206−22. DOI: 10.1128/aac.01206−22.
[31] Ruiz C, Levy SB. Many chromosomal genes modulate MarA-mediated multidrug resistance in *Escherichia coli*[J]. *Antimicrob Agents Chemother*, 2010, 54(5): 2125−2134. DOI: 10.1128/aac.01420−09.
[32] Li SS, Wu J, Ma N, et al. Prediction of genome-wide imipenem resistance features in *Klebsiella pneumoniae* using machine learning[J]. *J Med Microbiol*, 2023, 72(2): 001657. DOI: 10.1099/jmm.0.001657.

王瑜昊 ORCID：0009−0009−7849−0058

作者贡献：

王瑜昊：数据收集、整理、分析、文章撰写
赵俊岭：文章设计、数据分析、文章撰写
黄佳：数据分析、撰写与指导
吴鑫淼：数据分析、整理
卢昕：数据分析、整理、撰写与指导
阚飙、李臻鹏：文章设计、撰写与指导

本文创新点和学术评论句见开放科学（OSID）平台，欢迎扫描开放科学（OSID）二维码，与作者开展交流互动

引用本文：王瑜昊，赵俊岭，黄佳，等. 基于泛基因组特征的机器学习模型预测肺炎克雷伯菌对美罗培南的表型耐药[J]. 疾病监测，2025，40(5)：653 − 659. DOI: 10.3784/jbjc.202411080633

Wang YH，Zhao JL，Huang J，et al. Establishment of machine learning models based on pan-genome features for prediction of phenotypic resistance of *Klebsiella pneumoniae* to meropenem[J]. *Dis Surveill*, 2025, 40(5): 653 − 659. DOI: 10.3784/jbjc.202411080633

（本文编辑：杨小平）