

基于全基因组测序技术筛选儿童金黄色葡萄球菌的致病相关分子标志物

陈建余¹ 王旭麟¹ 李文宇¹ 陈敏琪¹ 周俊立¹ 姚振江¹ 傅锦坚² 叶小华¹

(1. 广东药科大学公共卫生学院, 广东广州 510310; 2. 柳州市妇幼保健院, 广西柳州 545001)

[摘要] **目的** 探究儿童金黄色葡萄球菌(以下简称“金葡萄”)的分子特征, 并比较不同类型菌株(感染菌株与定植菌株)的分子特征差异性, 从而揭示金葡萄致病相关分子标志物。**方法** 采用横断面研究设计, 对社区健康儿童进行鼻咽拭子采样和医院感染儿童进行临床标本采样。采用全基因组测序技术检测金葡萄耐药基因和毒素基因。采用随机森林方法筛选金葡萄致病相关标志物。**结果** 共检出512株金葡萄(272株感染菌株、240株定植菌株)。对于毒素基因, 感染菌株组的肠毒素基因(*seb*和*sep*)、胞外酶编码基因(*splA*、*splB*、*splE*和*edinC*)、杀白细胞毒素基因(*lukD*、*lukE*、*lukF-PV*和*lukS-PV*)和表皮剥脱毒素基因(*eta*和*etb*)携带率均高于定植菌株组(均 $P<0.05$), 但肠毒素基因(*sec*、*sec3*、*seg*、*seh*、*sei*、*sel*、*sem*、*sen*、*seo*和*seu*)携带率低于定植菌株组(均 $P<0.05$)。对于耐药基因, 感染菌株组的*lnuA*、*lnuG*、*aadD*、*tetK*和*dfcG*耐药基因携带率明显高于定植菌株组(均 $P<0.05$)。随机森林模型筛选金葡萄致病相关标志物, 筛选前后模型交叉验证的正确率分别为69%、68%, 曲线下面积分别为0.75、0.70; 最终筛选出16个致病相关标志物 [*sem*、*etb*、*splE*、*sep*、*ser*、*mecA*、*lnuA*、*sea*、*blaZ*、*cat(pC233)*、*blaTEM-1A*、*aph(3')-III*、*ermB*、*ermA*、*ant(9)-Ia*和*ant(6)-Ia*]; 变量重要性排序中前5个最重要变量为*sem* ($OR=0.40$)、*etb* ($OR=3.95$)、*splE* ($OR=1.68$)、*sep* ($OR=3.97$)、*ser* ($OR=1.68$)。**结论** 随机森林模型可筛选出金葡萄致病相关标志物且模型预测效果较优, 为追溯高致病性金葡萄和开展精准的靶向干预提供遗传学证据。
[中国当代儿科杂志, 2023, 25 (11): 1161-1169]

[关键词] 金黄色葡萄球菌; 分子特征; 随机森林; 全基因组测序; 儿童

Screening of pathogenic molecular markers of *Staphylococcus aureus* in children based on whole genome sequencing technology

CHEN Jian-Yu, WANG Xu-Lin, LI Wen-Yu, CHEN Min-Qi, ZHOU Jun-Li, YAO Zhen-Jiang, FU Jin-Jian, YE Xiao-Hua. School of Public Health, Guangdong Pharmaceutical University, Guangzhou 510310, China (Ye X-H, Email: smalltomato@163.com)

Abstract: Objective To explore the molecular characteristics of *Staphylococcus aureus* (*S. aureus*) in children, and to compare the molecular characteristics of different types of strains (infection and colonization strains) so as to reveal pathogenic molecular markers of *S. aureus*. **Methods** A cross-sectional study design was used to conduct nasopharyngeal swab sampling from healthy children in the community and clinical samples from infected children in the hospital. Whole genome sequencing was used to detect antibiotic resistance genes and virulence genes. A random forest method was used to screen pathogenic markers. **Results** A total of 512 *S. aureus* strains were detected, including 272 infection strains and 240 colonization strains. For virulence genes, the carrying rates of enterotoxin genes (*seb* and *sep*), extracellular enzyme coding genes (*splA*, *splB*, *splE* and *edinC*), leukocytotoxin genes (*lukD*, *lukE*, *lukF-PV* and *lukS-PV*) and epidermal exfoliating genes (*eta* and *etb*) in infection strains were higher than those in colonization strains. But the carrying rates of enterotoxin genes (*sec*, *sec3*, *seg*, *seh*, *sei*, *sel*, *sem*, *sen*, *seo* and *seu*) were lower in infection strains than in colonization strains ($P<0.05$). For antibiotic resistance genes, the carrying rates of *lnuA*, *lnuG*, *aadD*, *tetK*

[收稿日期] 2023-04-28; [接受日期] 2023-10-08

[基金项目] 广东省基础与应用基础研究基金(2023A1515011583); 国家自然科学基金(81973069)。

[作者简介] 陈建余, 男, 硕士研究生。

[通信作者] 叶小华, 女, 教授。Email: smalltomato@163.com。

and *dfrG* were significantly higher in infection strains than in colonization strains ($P<0.05$). The accuracy of cross-validation of the random forest model for screening pathogenic markers of *S. aureus* before and after screening was 69% and 68%, respectively, and the area under the curve was 0.75 and 0.70, respectively. The random forest model finally screened out 16 pathogenic markers (*sem*, *etb*, *splE*, *sep*, *ser*, *mecA*, *lnuA*, *sea*, *blaZ*, *cat(pC233)*, *blaTEM-1A*, *aph(3')-III*, *ermB*, *ermA*, *ant(9)-Ia* and *ant(6)-Ia*). The top five variables in the variable importance ranking were *sem* ($OR=0.40$), *etb* ($OR=3.95$), *splE* ($OR=1.68$), *sep* ($OR=3.97$), and *ser* ($OR=1.68$). **Conclusions** The random forest model can screen out pathogenic markers of *S. aureus* and exhibits a superior predictive performance, providing genetic evidence for tracing highly pathogenic *S. aureus* and conducting precise targeted interventions.

[Chinese Journal of Contemporary Pediatrics, 2023, 25(11): 1161-1169]

Key words: *Staphylococcus aureus*; Molecular characteristic; Random forest; Genome wide sequencing; Child

金黄色葡萄球菌 (*Staphylococcus aureus*, 以下简称“金葡菌”) 是儿童感染的重要致病菌之一^[1]。金葡菌的多重耐药率较高, 且社区和医院中的金葡菌交叉传播风险较高, 对人类健康造成极大威胁, 给全球带来了沉重的疾病经济负担, 已成为国际上日益关注的临床问题和公共卫生问题^[1]。既往研究表明, 儿童人群中的金葡菌鼻腔定植现象较为严重且明显高于成人, 在婴幼儿时期的定植率高达 45%^[2]; 同时, 儿童身体各方面尚未发育完善, 对病原体的抵抗力较低, 儿童金葡菌感染率明显高于成人, 且一旦感染则更容易发展为菌血症、骨髓炎、坏死性肺炎等侵袭性疾病^[3], 提示儿童是金葡菌定植和感染的重点高危人群。既往研究主要局限于研究医院感染人群金葡菌的耐药性、分子特征, 而关于健康儿童鼻腔定植金葡菌的分子流行病学研究仍缺乏系统报告, 且金葡菌致病相关标志物仍未阐明。金葡菌感染菌株与定植菌株的分子分型具有一定的相似性, 仅通过克隆群、多位点序列分型等传统分子分型方法难以区分菌株间的微小遗传变异, 有必要通过高通量全基因组测序技术阐明两类金葡菌间全基因组水平上的潜在遗传差异。金葡菌的致病力强弱与其自身含有的毒力因子高度相关, 其中与金葡菌感染有关的毒力因子包括肠毒素、溶血素、中毒休克综合征毒素、杀白细胞毒素等。2020 年研究^[4]表明, 中毒休克综合征毒素基因 *tsst-1*、表皮剥脱毒素基因 *eta* 和 *etb* 的表达可增强金葡菌的毒力及致病力, *fnbA*、*cna*、*sdrE*、*sej*、*eta*、*hlg* 和 *ica* 基因在感染菌株中比定植菌株更加普遍, 因此这些毒力因子为揭示金葡菌的致病相关标志物提供重要思路。随机森林是一种基于决策树的非参数模型, 可以有效处理基因组数据的高维、高相关性等问题, 其特点是可以对变量的重要性进行排序和识别致病相关因子, 为揭示金葡菌致病相关标志物提供新策略^[5]。本研究对 512 株金葡菌

(272 株感染菌株和 240 株定植菌株) 进行全基因组测序分析, 全面检测金葡菌的耐药基因和毒素基因, 运用随机森林筛选金葡菌致病相关分子标志物, 为追溯高致病性金葡菌和开展精准的靶向干预提供遗传学证据。

1 资料与方法

1.1 研究对象

本研究采用横断面研究设计。(1) 感染菌株来源: 收集 2014—2017 年在柳州市妇幼保健院就诊的感染儿童临床标本。感染儿童的纳入标准: 年龄 ≤ 7 岁; 家长签署知情同意书; 有临床感染症状 (如咳嗽、听诊异常、呼吸困难、发热超过 38℃ 等), 且从感染部位 (包括痰液、脓液、血液、支气管肺泡灌洗液等) 采集的临床样本中分离和鉴定出符合标准的阳性金葡菌, 最终确定为金葡菌感染的患儿。排除标准: 同一患儿的不同部位采集标本, 取主要发病部位的标本; 同一患儿对于同一疾病多次采样时, 只取首次标本。感染儿童的信息采集包括: 年龄、性别、病案号等基本信息; 标本来源、采样时间、疾病诊断等临床信息。(2) 定植菌株来源: 采用分层整群抽样抽取 2018 年 4—6 月期间广西柳州市 6 所幼儿园, 开展健康幼儿园儿童的鼻咽拭子采样。健康儿童的纳入标准: 年龄 ≤ 7 岁的健康儿童, 且家长签署知情同意书。排除标准: 调查前 7 d 内患有肺炎、支气管炎等急性感染性疾病的儿童, 以及患有血友病等易出血疾病导致不便采样的儿童。共调查 1 702 名健康儿童, 检出 277 株金葡菌, 其中 240 株成功进行全基因组测序分析。本研究获得柳州市妇幼保健院医学伦理委员会 (批件号: 2018-085) 和广东药科大学伦理委员会 (批件号: 2015-20) 批准。

1.2 菌株分离鉴定

菌株分离鉴定试验包括：甘露醇氯化钠琼脂培养基显色反应、革兰染色、纯化培养、溶血试验及血浆凝固酶试验，依次进行上述试验，结果阳性则鉴定为金葡菌。

1.3 全基因组测序分析

经菌液制备、试剂盒提取DNA、质量控制后获得金葡菌DNA，经分光光度检测仪检测浓度和纯度及格的基因组DNA进行全基因组测序，采用Illumina HiSeq 2000测序仪对基因组DNA进行双端测序，使用FastQC进行原始测序数据的质控^[6]，采用SPAdes 3.12.0软件进行序列拼接。使用基因组流行病学中心（Center for Genomic Epidemiology, CGE）网站（<http://www.genomicepidemiology.org/>）中的ResFinder 4.1（<https://cge.cbs.dtu.dk/services/ResFinder/>）进行序列比对，对15类2 208个耐药基因进行检测^[7]。使用CGE中VirulenceFinder 2.0（<https://cge.cbs.dtu.dk/services/VirulenceFinder/>）进行序列比对获得菌株的毒素基因，主要包括免疫逃逸基因、胞外酶基因、溶血毒素基因、肠毒素基因、表皮剥脱素基因、杀白细胞毒素基因、中毒休克综合征毒素基因。以*S. aureus* subsp. *aureus* MRSA252（accession number: NC_002952）为参考基因组，使用Snippy软件（<https://github.com/tseemann/snippy/>）对菌株进行核心单核苷酸多态性鉴定，使用Gubbins软件去除基因重组或水平基因转移相关区域^[8]。使用Fasttree软件构建基于核心基因组的系统进化树，采用广义时间可逆替代（general time reversible, GTR）模型^[9]，估计方法为极大似然估计法，Bootstrap值设置为100。使用Chiplot网站（<http://www.evolgenius.info/>）对系统进化树进行可视化和美化。

1.4 统计学分析

采用EpiData 3.0软件进行数据的双人双录入，并通过一致性检验保证数据的准确性。采用Stata 16.0和R 4.1.2软件对数据进行统计学分析。计数资料以频数和百分率（%）表示，采用Pearson χ^2 检验或Fisher确切概率法比较金葡菌感染菌株与定植菌株的耐药基因、毒素基因携带率，从而初步筛选致病相关因子。 $P<0.05$ 为差异有统计学意义。采用单因素logistic回归或确切logistic回归模型估计OR及95%CI。运用机器学习方法中的随机森林

进一步筛选重要的致病相关特征变量，模型以菌株类型（1=感染菌株，0=定植菌株）作为结局变量，以分子特征（毒素基因和耐药基因）作为自变量。随机森林分析中，通过randomForestSRC程序包中随机森林的变量捕获法（variable selection using random forests, VSURF）对重要特征变量进行筛选^[10]；采用randomForest程序包对最终模型中的变量进行重要性评分，特征变量的重要性以平均基尼指数减少量（mean decrease in the Gini, MDG）排序，从而判断各个变量在模型中的重要性。采用组内回代和十折交叉验证方法评价最终模型的预测效果，评价指标包括正确率、灵敏度、特异度、阳性预测值、阴性预测值、受试者操作特征曲线（receiver operating characteristic curve, ROC曲线）。

2 结果

2.1 基本情况

有272株金葡菌（即感染菌株组）来源于医院感染儿童，年龄为0~7岁，中位数为2个月，172株（63.2%）来源于男性、100株（36.8%）来源于女性。有240株鼻腔定植金葡菌（即定植菌株组）来源于健康儿童，年龄为3~7岁，中位数为5岁，124株（51.7%）来源于男性、116株（48.3%）来源于女性。

2.2 定植菌株组与感染菌株组的毒素基因携带情况

金葡菌感染菌株组的*seb*、*sep*、*splA*、*splB*、*splE*、*edinC*、*lukD*、*lukE*、*lukF-PV*、*lukS-PV*、*eta*和*etb*基因携带率高于定植菌株组（均 $P<0.05$ ）；但感染菌株组的*sec*、*sec3*、*seg*、*seh*、*sei*、*sel*、*sem*、*sen*、*seo*和*seu*基因携带率低于定植菌株组（均 $P<0.05$ ），其他毒素基因携带率比较差异无统计学意义（ $P>0.05$ ）。见表1。

2.3 定植菌株组与感染菌株组的耐药基因携带情况

金葡菌感染菌株组的*lnuA*、*aadD*、*tetK*和*dfrG*基因携带率明显高于定植菌株组（ $P<0.05$ ）。但感染菌株组的*blaTEM-1A*基因携带率低于定植菌株（ $P<0.05$ ）。其余耐药基因携带率比较差异均无统计学意义（ $P>0.05$ ）。见表2。

表 1 不同类型金葡萄菌的毒素基因携带情况比较 [株 (%)]

毒素基因	定植菌株组 (n=240)	感染菌株组 (n=272)	χ^2 值	P 值	OR(95%CI)
肠毒素基因					
<i>sea</i>	35(14.6)	45(16.5)	0.37	0.542	1.16(0.72~1.88)
<i>seb</i>	40(16.7)	66(24.3)	4.48	0.034	1.60(1.03~2.48)
<i>sec</i>	67(27.9)	46(16.9)	8.98	0.003	0.53(0.34~0.80)
<i>sec3</i>	50(20.8)	15(5.5)	26.99	<0.001	0.22(0.12~0.41)
<i>sed</i>	20(8.3)	34(12.5)	2.35	0.126	1.57(0.87~2.81)
<i>seg</i>	123(51.3)	76(27.9)	29.15	<0.001	0.37(0.26~0.53)
<i>seh</i>	28(11.7)	16(5.9)	5.43	0.020	0.47(0.25~0.90)
<i>sei</i>	125(52.1)	77(28.3)	30.17	<0.001	0.36(0.25~0.52)
<i>sej</i>	20(8.3)	36(13.2)	3.15	0.076	1.68(0.94~2.99)
<i>sek</i>	39(16.3)	59(21.7)	2.44	0.118	1.42(0.91~2.23)
<i>sel</i>	68(28.3)	48(17.6)	8.31	0.004	0.54(0.36~0.82)
<i>sem</i>	127(52.9)	85(31.3)	24.67	<0.001	0.40(0.28~0.58)
<i>sen</i>	125(52.1)	76(27.9)	31.16	<0.001	0.36(0.25~0.51)
<i>seo</i>	127(52.9)	84(30.9)	25.55	<0.001	0.40(0.28~0.57)
<i>sep</i>	12(5.0)	47(17.3)	18.56	<0.001	3.97(2.05~7.68)
<i>seq</i>	39(16.3)	60(22.1)	2.76	0.097	1.46(0.93~2.28)
<i>ser</i>	20(8.3)	36(13.2)	3.15	0.076	1.68(0.94~2.99)
<i>seu</i>	124(51.7)	76(27.9)	30.15	<0.001	0.36(0.25~0.52)
免疫逃逸基因					
<i>sak</i>	196(81.7)	210(77.2)	1.55	0.214	0.76(0.49~1.17)
<i>scn</i>	233(97.1)	264(97.1)	<0.01	0.987	0.99(0.35~2.78)
胞外酶编码基因					
<i>aur</i>	237(98.8)	268(98.5)	—	1.000 ^a	0.84(0.19~3.83)
<i>splA</i>	119(49.6)	188(69.1)	20.27	<0.001	2.28(1.59~3.26)
<i>splB</i>	118(49.2)	191(70.2)	23.62	<0.001	2.44(1.70~3.50)
<i>splE</i>	84(35.0)	129(47.4)	8.10	0.004	1.68(1.17~2.39)
<i>edinA</i>	6(2.5)	3(1.1)	1.44	0.230	0.43(0.11~1.76)
<i>edinB</i>	3(1.3)	10(3.7)	3.03	0.082	3.02(0.82~11.1)
<i>edinC</i>	6(2.5)	24(8.8)	9.24	0.002	3.77(1.52~9.40)
杀白细胞毒素基因					
<i>lukD</i>	121(50.4)	191(70.2)	21.01	<0.001	2.32(1.61~3.33)
<i>lukE</i>	121(50.4)	193(71.0)	22.68	<0.001	2.40(1.67~3.46)
<i>lukF-PV</i>	30(12.5)	59(21.7)	7.50	0.006	1.94(1.20~3.13)
<i>lukS-PV</i>	30(12.5)	60(22.1)	8.04	0.005	1.98(1.23~3.20)
溶血毒素基因					
<i>hlgA</i>	239(99.6)	272(100)	—	0.469 ^a	1.13(0.03~∞) ^b
<i>hlgB</i>	239(99.6)	269(98.9)	—	0.775 ^a	0.38(0.01~4.72)
<i>hlgC</i>	240(100)	272(100)	—	—	—
表皮剥脱毒素基因					
<i>eta</i>	7(2.9)	20(7.4)	5.02	0.025	2.64(1.10~6.36)
<i>etb</i>	6(2.5)	25(9.2)	10.04	0.002	3.95(1.59~9.79)
中毒性休克综合征毒素基因					
<i>tsst-1</i>	21(8.8)	14(5.1)	2.60	0.107	0.57(0.28~1.14)

注：^a使用 Fisher 确切概率法；^b采用确切 logistic 回归。

表2 不同类型金葡萄菌的耐药基因携带情况比较 [株 (%)]

耐药基因	定植菌株组 (n=240)	感染菌株组 (n=272)	χ^2 值	P 值	OR(95% CI)
大环内酯类					
<i>erm(C)</i>	167(69.6)	198(72.8)	0.64	0.423	1.16(0.80~1.72)
<i>ermA</i>	7(2.9)	5(1.8)	0.64	0.421	0.62(0.15~2.32)
<i>ermB</i>	51(21.3)	59(21.7)	0.01	0.903	1.02(0.67~1.55)
<i>erm(T)</i>	1(0.4)	0(0)	–	0.469 ^a	0.88(0.00~34.41) ^b
<i>msr(A)</i>	2(0.8)	0(0)	–	0.219 ^a	0.36(0.00~4.69) ^b
<i>mph(C)</i>	1(0.4)	0(0)	–	0.469 ^a	0.88(0.00~34.41) ^b
<i>lnuA</i>	20(8.3)	50(18.4)	10.90	0.001	2.48(1.43~4.30)
<i>lnuG</i>	7(2.9)	18(6.6)	3.76	0.052	2.36(0.97~5.75)
<i>vga(A)LC</i>	0(0)	1(0.4)	–	1.000 ^a	0.88(0.02~∞) ^b
氨基糖苷类					
<i>aac(6)-aph(2'')</i>	0(0)	3(1.1)	–	0.251 ^a	3.41(0.37~∞) ^b
<i>aadD</i>	19(7.9)	55(20.2)	15.61	<0.001	2.95(1.69~5.13)
<i>ant(6)-I a</i>	40(16.7)	50(18.4)	0.26	0.611	1.13(0.71~1.78)
<i>ant(9)-I a</i>	7(2.9)	3(1.1)	2.19	0.139	0.37(0.09~1.45)
<i>aph(2'')-I a</i>	10(4.2)	11(4.0)	0.01	0.944	0.97(0.40~2.32)
<i>aph(3')-II a</i>	1(0.4)	0(0)	–	0.469 ^a	0.88(0.00~34.41) ^b
<i>aph(3')-III</i>	44(18.3)	50(18.4)	<0.001	0.989	1.00(0.64~1.57)
四环素类					
<i>tetK</i>	75(31.3)	135(49.6)	17.81	<0.001	2.17(1.51~3.11)
<i>tet(L)</i>	0(0)	1(0.4)	–	1.000 ^a	0.88(0.02~∞) ^b
<i>tet(M)</i>	2(0.8)	0(0)	–	0.219 ^a	0.36(0.00~0.47) ^b
β-内酰胺类					
<i>blaZ</i>	229(95.4)	257(94.5)	0.23	0.632	0.82(0.37~1.83)
<i>mecA</i>	81(33.8)	78(28.7)	1.53	0.216	0.79(0.54~1.15)
<i>blaTEM-116</i>	1(0.4)	0(0)	–	0.469 ^a	0.88(0.00~34.41) ^b
<i>blaTEM-1A</i>	6(2.5)	0(0)	–	0.010 ^a	0.11(0.00~0.74) ^b
氯霉素类					
<i>cat</i>	2(0.8)	3(1.1)	–	1.000 ^a	1.32(0.15~16.01)
<i>catA1</i>	3(1.3)	0(0)	–	0.102 ^a	0.23(0.00~2.13) ^b
<i>cat(pC221)</i>	0(0)	5(1.8)	–	0.064 ^a	6.01(0.81~∞) ^b
<i>cat(pC233)</i>	15(6.3)	30(11.0)	3.63	0.057	1.86(0.97~3.54)
甲氧苄啶类					
<i>dfzG</i>	7(2.9)	23(8.5)	7.09	0.008	3.07(1.29~7.30)
<i>dfzA31</i>	0(0)	1(0.4)	–	1.000 ^a	0.88(0.02~∞) ^b
磷霉素类					
<i>fosB</i>	3(1.3)	0(0)	–	0.102 ^a	0.23(0.00~2.13) ^b
夫西地酸类					
<i>fusB</i>	0(0)	2(0.7)	–	0.501 ^a	2.13(0.17~∞) ^b
耐消毒剂基因					
<i>qacA</i>	0(0)	2(0.7)	–	0.501 ^a	2.13(0.16~∞) ^b

注：^a使用 Fisher 确切概率法；^b采用确切 logistic 回归。

2.4 随机森林筛选金葡萄菌致病相关分子特征

基于核心单核苷酸多态性的系统进化树（图 1）表明：感染菌株与定植菌株在系统进化树上的

分布没有明显差异，提示两者可能具有相似的遗传背景，需进一步揭示致病相关分子特征。以菌株类型（1=感染菌株，0=定植菌株）为因变量，

建立随机森林模型筛选儿童金葡菌致病相关的重要特征变量 (69 个), 采用 VSURF 法最终筛选出 16 个特征变量。对变量筛选前后的模型进行预测效果评价, 结果表明: 变量筛选前后预测模型的交叉验证正确率分别为 69%、68% (表 3); 变量筛选前模型的曲线下面积 (area under the curve, AUC) 稍高于筛选后 (0.75 vs 0.70) (图 2A~B)。最终筛选出的 16 个特征变量为肠毒素基因 (*sem*、*sep*、*ser*、*sea*)、表皮剥脱毒素基因 *etb*、胞外酶编码基因 *splE*、 β -内酰胺类耐药基因 (*mecA*、*blaZ*、*blaTEM-1A*)、大环内酯类耐药基因 (*ermA*、*ermB*、

lnuA)、氯霉素类耐药基因 *cat(pC233)*、氨基糖苷类耐药基因 [*aph(3')-III*、*ant(9)-Ia*、*ant(6)-Ia*]; 变量重要性排序结果显示, 前 5 个最重要的特征变量分别为 *sem*、*etb*、*splE*、*sep*、*ser* (图 2C)。使用随机森林模型进一步对 16 个重要致病相关标志物进行风险预测, 结果显示: 有 10 个致病相关标志物位于对角线上方, 提示携带这些致病相关标志物会增加金葡菌的致病风险; 而有 6 个致病相关标志物位于对角线下方, 提示携带这些致病标志物会降低金葡菌的致病风险 (图 2D)。

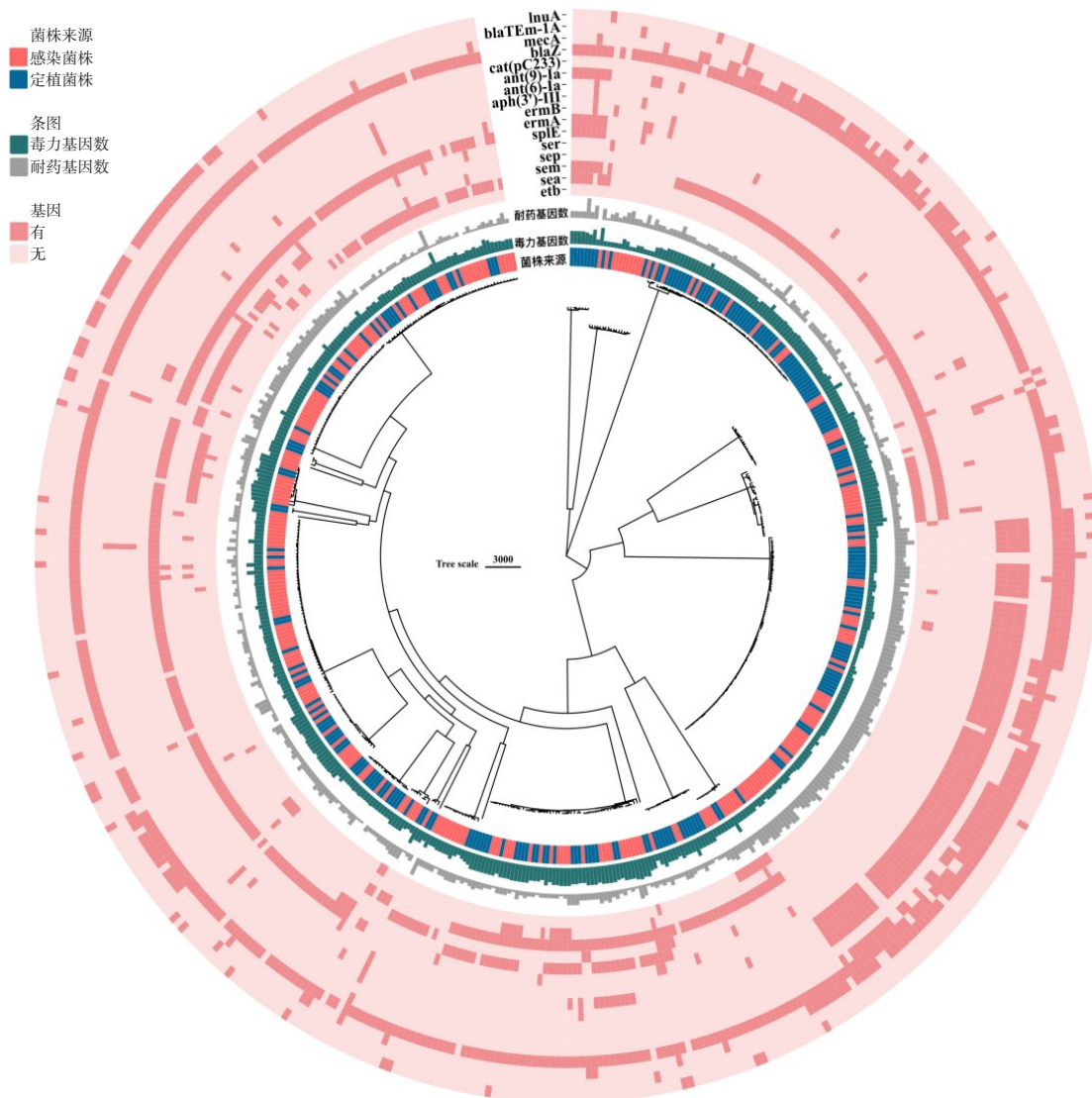


图1 512株金葡萄菌的系统进化树 从内环到外环的变量分别为菌株来源 (红色为感染菌株, 蓝色为定植菌株)、毒力基因数、耐药基因数、16个致病相关特征变量 (深红表示有, 浅红表示无)。

表 3 基于分子特征的随机森林模型预测效果评价

评价指标	筛选前模型（69 个）		筛选后模型（16 个）	
	组内回代	交叉验证	组内回代	交叉验证
正确率	0.81	0.69	0.74	0.68
灵敏度	0.78	0.68	0.71	0.66
特异度	0.86	0.71	0.79	0.72
阳性预测值	0.90	0.79	0.86	0.81
阴性预测值	0.71	0.57	0.59	0.53

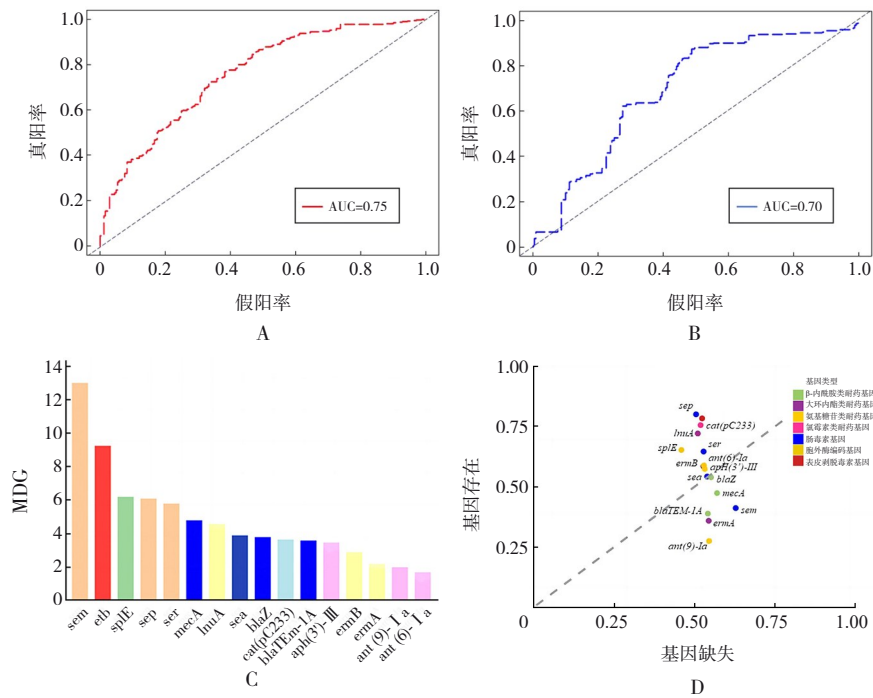


图 2 随机森林模型的拟合效果和风险得分图 A: 变量筛选前模型的 ROC 曲线。B: 变量筛选后模型的 ROC 曲线。C: 最终模型中 16 个特征变量的重要性排序。D: 最终模型中 16 个特征变量的风险得分图。

3 讨论

关于金葡菌的耐药机制研究较多，常见于以下几类耐药机制^[11-14]：（1）酶降解抗生素，如β-内酰胺酶；（2）产生拮抗物质，减少抗生素的有效浓度；（3）改变细胞靶标，使抗生素不能正常发挥作用；（4）主动外排泵作用。*blaZ*基因受 *blaR1*-*BlaI* 系统的调节控制，*blaR1* 蛋白受到β-内酰胺类抗生素的刺激后，导致抑制蛋白 *BlaI* 水解脱离结合位点，编码产生β-内酰胺酶，继而破坏β-内酰胺从而使金葡菌对青霉素类抗生素耐药^[11]。*mecA* 编码 PBP2a 蛋白，产生过量的 PBP 蛋白消耗β-内酰胺类药物，从而导致金葡菌对β-内酰胺类药物耐药。既往研究^[12]显示，虽然 *blaTEM* 基因编码的β-内酰胺酶主要通过丝氨酸残基灭活β-内酰胺环产生耐药性，但是 *blaTEM* 在多重耐药菌株和

敏感菌株中均被检出且差异无统计学意义，提示 *blaTEM* 可能不是导致其广泛耐药的主要原因；值得注意的是，本研究的随机森林预测模型揭示，*blaTEM-1A* 基因是金葡菌致病相关标志物，提示其可能影响菌株的致病能力。*erm* 编码核糖体甲基化酶催化金葡菌 23S rRNA 发生甲基化反应使核糖体靶位点发生变化，进而阻止大环内酯类药物与核糖体靶位结合，从而导致金葡菌对大环内酯类药物耐药^[13]。*lnuA* 基因编码核苷酸转移酶，从而介导对林可酰胺类抗生素的耐药性^[14]。

细菌的毒力因子在感染性疾病中起着关键作用，它是由多种毒素基因决定的。金葡菌具有数十种毒素因子，包括肠毒素、表皮剥脱毒素、胞外酶等。本研究随机森林结果显示，金葡菌致病相关的毒素基因有肠毒素基因（*sem*、*sep*、*ser* 和 *sea*）、表皮剥脱毒素基因（*eth*）、胞外酶编码基因

(*splE*)。葡萄球菌肠毒素在食源性中毒中最常见,它不仅可引起食物中毒,还可引起中毒性休克综合征等严重的侵袭性疾病^[15]。本研究中,感染菌株组 *sep* 基因的携带率明显高于定植菌株组,但是 *sem* 基因的携带率明显低于定植菌株组,提示感染菌株与定植菌株存在毒力差异,与中国台湾的一项研究结果一致^[16]。表皮剥脱毒素A和表皮剥脱毒素B是导致大多数人类葡萄球菌烫伤样皮肤综合征的重要原因^[17];本研究中,感染菌株组较定植菌株组的 *eta* (7.4% vs 2.9%)、*etb* (9.2% vs 2.5%) 基因携带率高,以上结果提示表皮剥脱毒素与菌株的致病能力密切相关。*splE* 编码丝氨酸蛋白酶样蛋白,可以在健康人,尤其是哮喘患者中诱导Ⅱ型超敏反应^[18];西班牙一项研究提示, *splE* 基因可作为金葡菌血症的致病相关标志物^[19]。以上结果提示不同类型疾病可能存在致病相关标志物的差异,因此后续研究可针对不同疾病类型进一步探讨潜在的的特异性标志物。

基因组数据具有“高维度、小样本”的数据特点,且存在“非线性、高相关”的复杂关联,使得传统分析方法不再适用,这对基因组关联分析提出了新挑战。随机森林是利用多棵决策树对样本进行训练和预测的机器学习算法,能有效处理线性、非线性、具有交互作用的复杂数据,具有抗噪声、防止过拟合、不受共线性影响的重要特征。随机森林不仅具有良好的预测能力,且具有精度高、稳定性好、易操作等优点,因此已成为处理医学领域基因组数据的精确机器学习方法之一。为了获得更准确的风险预测模型,本研究以金葡菌的常见致病相关基因(32个耐药基因和37个毒素基因)为预测变量、菌株类型(感染和定植菌株)为因变量建立随机森林模型,最终模型的16个预测因子分别为肠毒素基因(*sem*、*sep*、*ser*和*sea*)、表皮剥脱毒素基因*etb*、胞外酶编码基因*splE*、 β -内酰胺类耐药基因(*mecA*、*blaZ*和*blaTEM-1A*)、大环内酯类耐药基因(*ermA*、*ermB*和*lnuA*)、氯霉素类耐药基因*cat(pC233)*、氨基糖苷类耐药基因 [*aph(3')-III*、*ant(9)-I a*和*ant(6)-I a*],最终模型的交叉验证正确率为68%,AUC为0.70,提示模型的拟合效果较好。随机森林模型揭示:前5个最重要的预测变量为*sem*、*etb*、*splE*、*sep*、*ser*;其中*sep* (*OR*=3.97)、*etb* (*OR*=3.95)、*ser* (*OR*=1.68)和*splE* (*OR*=1.68)明显提高金葡菌的致病风险,提示肠毒素、表皮剥脱毒素和胞

外酶是金葡菌的重要致病相关标志物。以上结果揭示这些致病相关分子标志物有较大的潜力预测金葡菌致病株。

综上,本研究使用高通量全基因组测序技术全面检测金葡菌基因组的耐药基因和毒素基因,采用随机森林筛选出金葡菌的16个致病相关标志物(6个毒素基因和10个耐药基因),且模型预测效果较优,为追溯高致病性金葡菌和开展精准的靶向干预提供遗传学证据。本研究也存在一定局限性:(1)本研究提示致病相关特征变量与结局之间存在统计学关联,但筛选的致病相关标志物与疾病状态之间不一定存在因果关系,且仍未能深入阐明潜在的生物学功能与作用机制,因此今后可以结合生物学功能富集分析、通路富集分析对其潜在作用机制进行深入研究。(2)本研究于2014—2018年间对金葡菌感染患儿和健康儿童进行金葡菌采样,菌株的采集时间较早;但是,本研究主要从基因组水平上探究金葡菌的致病相关标志物,且既往研究表明菌株分子特征与菌株的采样时间无关^[20]。今后可进一步开展多中心调查来增大样本量和补充新菌株。

利益冲突声明:所有作者均声明不存在利益冲突。

[参 考 文 献]

- [1] Ahmad-Mansour N, Loubet P, Pouget C, et al. *Staphylococcus aureus* toxins: an update on their pathogenic properties and potential treatments[J]. *Toxins (Basel)*, 2021, 13(10): 677. PMID: 34678970. PMCID: PMC8540901. DOI: 10.3390/toxins13100677.
- [2] Wertheim HF, Melles DC, Vos MC, et al. The role of nasal carriage in *Staphylococcus aureus* infections[J]. *Lancet Infect Dis*, 2005, 5(12): 751-762. PMID: 16310147. DOI: 10.1016/S1473-3099(05)70295-4.
- [3] Hamdy RF, Dona D, Jacobs MB, et al. Risk factors for complications in children with *Staphylococcus aureus* bacteremia[J]. *J Pediatr*, 2019, 208: 214-220.e2. PMID: 30879729. DOI: 10.1016/j.jpeds.2018.12.002.
- [4] Shettigar K, Murali TS. Virulence factors and clonal diversity of *Staphylococcus aureus* in colonization and wound infection with emphasis on diabetic foot infection[J]. *Eur J Clin Microbiol Infect Dis*, 2020, 39(12): 2235-2246. PMID: 32683595. PMCID: PMC7669779. DOI: 10.1007/s10096-020-03984-8.
- [5] 宋欠欠,李轶群,侯艳,等.随机森林的变量捕获方法在高维数据变量筛选中的应用[J]. *中国卫生统计*, 2015, 32(1): 49-53.
- [6] Pribelski A, Antipov D, Meleshko D, et al. Using SPAdes de

- novo assembler*[J]. *Curr Protoc Bioinformatics*, 2020, 70(1): e102. PMID: 32559359. DOI: 10.1002/cpbi.102.
- [7] Bortolaia V, Kaas RS, Ruppe E, et al. ResFinder 4.0 for predictions of phenotypes from genotypes[J]. *J Antimicrob Chemother*, 2020, 75(12): 3491-3500. PMID: 32780112. PMCID: PMC7662176. DOI: 10.1093/jac/dkaa345.
- [8] Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins[J]. *Nucleic Acids Res*, 2015, 43(3): e15. PMID: 25414349. PMCID: PMC4330336. DOI: 10.1093/nar/gku1196.
- [9] Catanzaro D, Pesenti R, Milinkovitch MC. A non-linear optimization procedure to estimate distances and instantaneous substitution rate matrices under the GTR model[J]. *Bioinformatics*, 2006, 22(6): 708-715. PMID: 16397009. DOI: 10.1093/bioinformatics/btk001.
- [10] Ishwaran H, Kogalur UB, Gorodeski EZ, et al. High-dimensional variable selection for survival data[J]. *J Am Stat Assoc*, 2010, 105(489): 205-217. DOI: 10.1198/jasa.2009.tm08622.
- [11] Peacock SJ, Paterson GK. Mechanisms of methicillin resistance in *Staphylococcus aureus*[J]. *Annu Rev Biochem*, 2015, 84: 577-601. PMID: 26034890. DOI: 10.1146/annurev-biochem-060614-034516.
- [12] 江培涛, 方敏, 刘棵文, 等. 仪征地区鲍曼不动杆菌碳青霉烯类抗菌药耐药相关酶表型和基因型分析[J]. *南京医科大学学报(自然科学版)*, 2019, 39(1): 54-61. DOI: 10.7655/NYDXBNS20190110.
- [13] 罗曼, 刘倩. 食源性金黄色葡萄球菌耐药机制分析[J]. *中国病原生物学杂志*, 2022, 17(6): 685-688. DOI: 10.13350/j.cjpb.220613.
- [14] Mahfouz AA, Said HS, Elfeky SM, et al. Inhibition of erythromycin and erythromycin-induced resistance among *Staphylococcus aureus* clinical isolates[J]. *Antibiotics (Basel)*, 2023, 12(3): 503. PMID: 36978370. PMCID: PMC10044026. DOI: 10.3390/antibiotics12030503.
- [15] Etter D, Schelin J, Schuppler M, et al. Staphylococcal enterotoxin C: an update on SEC variants, their structure and properties, and their role in foodborne intoxications[J]. *Toxins (Basel)*, 2020, 12(9): 584. PMID: 32927913. PMCID: PMC7551944. DOI: 10.3390/toxins12090584.
- [16] Lo WT, Wang SR, Tseng MH, et al. Comparative molecular analysis of methicillin-resistant *Staphylococcus aureus* isolates from children with atopic dermatitis and healthy subjects in Taiwan[J]. *Br J Dermatol*, 2010, 162(5): 1110-1116. PMID: 20132206. DOI: 10.1111/j.1365-2133.2010.09679.x.
- [17] Oliveira D, Borges A, Simões M. *Staphylococcus aureus* toxins and their molecular activity in infectious diseases[J]. *Toxins (Basel)*, 2018, 10(6): 252. PMID: 29921792. PMCID: PMC6024779. DOI: 10.3390/toxins10060252.
- [18] Nordengrün M, Abdurrahman G, Treffon J, et al. Allergic reactions to serine protease-like proteins of *Staphylococcus aureus*[J]. *Front Immunol*, 2021, 12: 651060. PMID: 33833764. PMCID: PMC8021911. DOI: 10.3389/fimmu.2021.651060.
- [19] Pérez-Montarelo D, Viedma E, Larrosa N, et al. Molecular epidemiology of *Staphylococcus aureus* bacteremia: association of molecular factors with the source of infection[J]. *Front Microbiol*, 2018, 9: 2210. PMID: 30319561. PMCID: PMC6167439. DOI: 10.3389/fmicb.2018.02210.
- [20] Lange J, Heidenreich K, Higelin K, et al. *Staphylococcus aureus* pathogenicity in cystic fibrosis patients: results from an observational prospective multicenter study concerning virulence genes, phylogeny, and gene plasticity[J]. *Toxins (Basel)*, 2020, 12(5): 279. PMID: 32357453. PMCID: PMC7290773. DOI: 10.3390/toxins12050279.

(本文编辑: 杨丹)

(版权所有©2023 中国当代儿科杂志)